SHORT WAVE WIRELESS COMMUNICATION

including ultra-short waves

by

A. W. LADNER
A.M.I.C.E.

Formerly Principal of the Marconi School of Wireless Communication and Superintendent of Instruction, Marconi's Wireless Telegraph Co. Ltd.

and

C. R. STONER
B.Sc.(Eng.), M.I.E.E., M.I.R.E.

Reader in Electrical Engineering, Queen Mary College (University of London), formerly of Marconi's Wireless Telegraph Co. Ltd.

LONDON
CHAPMAN & HALL LTD.
37, ESSEX STREET, W.C.2

First published, 1923
Reprinted, 1934
Reprinted, 1935
Reprinted, 1935

PRINTED IN GREAT BRITAIN BY THE WHITEFRIARS PRESS LTD.
LONDON AND TONBRIDGE

BOUND BY G. AND J. KITCAT, LIMITED, LONDON
Flexiback Binding

PREFACE TO FIRST EDITION

THE discovery, in the third decade of the present century, of the utility of "short" waves for world-wide communication, in which amateur workers can claim a large share of the credit, produced a revolutionary change in the science of wireless and had a profound influence, both technical and economic, on world communication.

The introduction of short wave beam methods enabled wireless, which had up to this time only carried a very small percentage of the world's long distance communications, to offer a high-speed telegraph service in most ways vastly superior to that of the older methods and at a very much smaller capital cost. In addition, it provided the only means so far developed for commercial trans-oceanic telephony.

In presenting a book on the principles of short wave wireless communication, it is the aim of the authors to fill an obvious gap in current literature and supply a text-book which shall satisfy the needs not only of engineers and telegraphists engaged in wireless, but cater for the scientific amateur and those who have already an outline knowledge of long wave working, such as may be gained by reading any elementary text-book on the subject.

Although the book deals especially with short wave communication, the field has not been restricted entirely to the peculiarities of short waves. A self-contained treatise has been aimed at, principles common both to long and short waves being introduced where the matter is necessary for the clearer explanation of the main subject.

The authors wish to pay tribute to the genius of C. S. Franklin, who from the earliest days of wireless has contributed so much to its progress in every field of action, and has played such a great part in the successful utilisation of short waves. His brilliant work in the development of the Beam and guiding it to success should alone be sufficient to ensure him an honoured place in the history of communication engineering, and this represents but a small fraction of his contribution to the art.

In a book of this nature it is not possible to show adequate appreciation to our numerous friends who have supplied information, but we wish to make the following acknowledgments.

Firstly to Mr. Andrew Gray, until recently Technical General Manager of Marconi's Wireless Telegraph Co. Ltd., but for whose encouragement and kindly interest this book would not have been written. To Mr. T. L. Eckersley, who has been most helpful, and whose researches have supplied the bulk of the material for Chapter IV. To Mr. N. Wells, who has kindly checked the chapters on Transmitters, Feeders, Modulation and Aerials, and made many helpful suggestions. To Mr. Norman C. Stamford, who has assisted generally with the preparation of the book, and in the careful checking of the final manuscript.

To Marconi's Wireless Telegraph Co. Ltd., who have supplied much information and permitted the publication of the same, and to Mr. H. M. Dowsett for his light use of the Censor's pencil. Certain portions of the chapters on Modulation and Polar Diagrams of Aerial Arrays have previously appeared in the *Marconi Review*, and our thanks are due to the Editor for permission to republish. Thanks are also due for the loan of several blocks.

To the Imperial and International Communications Co. Ltd., for permission to publish information concerning their stations; particularly to Mr. J. Brown, Engineer-in-Charge of the Somerton Station, Mr. P. J. Woodward for information on the swinging of beams, and Mr. R. Keen for information regarding frequency measuring equipment.

To Standard Telephones and Cables Ltd., for information supplied on their transmitters, array systems and receivers.

To the Institute of Radio Engineers (U.S.A.) for permission to make use of papers which have appeared in their Journal, and to Messrs. Beverage and Carter, of the Radio Corporation of America, for checking the description of R.C.A. apparatus.

Our special thanks are due to Miss E. M. Elliot and Miss D. Turner for their great care in the preparation of the line drawings which illustrate the work.

1932.

CONTENTS

CHAPTER I

INTRODUCTION

SINCE the publication of the first edition of this book, the strangeness of working with short waves has worn off and they are accepted, not only as a simple and efficient means of wireless communication but, indeed, as the principal means. Although the advent of short waves is now past history, a very brief survey of the events leading up to their success may be of interest to the new reader, as probably in no branch of applied science has there been a more surprising " kink " in the ordinary line of progress.

Commercial wireless commenced at the beginning of the century and in the next two decades short, medium and long distance communication developed along orthodox lines. Short and medium waves appeared best suited to short and medium distances and long waves appeared to be necessary for long distances. The development of a number of very high-power stations capable of world-wide communication, in the early 1920's, was the culmination of a gradual logical evolution drawing inexorably to a long foreseen climax. That such stations had severe limitations was always evident, on account of the few routes on which they could operate satisfactorily because of atmospheric disturbance and also because their low-damped circuits limited the types of traffic that could be handled.

Thus the discovery, by amateurs, in 1921, of the possibilities of waves between 100 and 200 metres as a means of long-distance signalling, stimulated a world-wide and intensive research by a number of wireless organisations into their possibilities for serious work. In America, and on the Continent, transmitters and receivers without beam aerials were set up because there was a strong feeling prevalent that any beam effect would disappear when waves passed through the ionosphere, and beam aerials were very costly and largely unknown. In this country, however, Marconi put his faith in

the great magnifying effect of the Beam, and boldly commenced the erection of a chain of Beam Stations in this country and the Colonies. His optimism was fully justified as, within the space of four years (the time required to develop and build the first of these stations), the Marconi Beam produced results which greatly surpassed anything that had been obtained by any other means. Other countries were still obtaining unreliable results and using short waves only for auxiliary stations. Thus was fulfilled the dream of an Empire Chain of Wireless Stations, projected so many years before, but never accomplished.

The success of the Beam has had a profound influence on the history of communication. As far as Great Britain is concerned the Beam was an engineering triumph but an economic tragedy. It was a triumph technically because a wireless communication system was suddenly produced which could work direct to the furthest stations on the earth's surface, over routes hitherto unworkable by wireless means. The system was capable of conveying not only telegraphy at the highest possible speeds but also other forms of intelligence requiring a wide frequency-band.

Economically, the Beam was to prove something of a tragedy for Great Britain. When its capabilities were fully realised and adopted by other countries, Britain lost what she had had up till then, a virtual monopoly of handling world communications through her cable network. This she had obtained chiefly because of the geography of the British Commonwealth which required long trans-oceanic communications, and to the fortuitous circumstance that the Empire could provide so many lands and scattered islands dotted over the surface of the Earth on which the necessary repeater stations could be located. The cost of a wireless beam system is only a fraction of that of a cable, and in the passing years countries which hitherto made extensive use of the British cable routes set up direct beam services of their own, and much traffic was thereby diverted. But since the cable system has many advantages for telegraphy and therefore still forms a valuable national asset, a Government-sponsored merger was formed of all overseas cable and wireless telegraph services in order that the two systems might become complementary.

More recently still this Cable and Wireless Company has been nationalised so that all telegraph and telephone services are now Government-owned and controlled.

An essential link in any radio communication is the path between transmitter and receiver. Since its properties are complex, selective and variable, it may be helpful to consider a broad, if arbitrary, division of the radio wave-range into groups as shown by the Table below, in order to show how short and ultra-short waves fit into the general communication picture.

TABLE I

Classification of Radio Waves.

Classification.	Frequencies. *Kilocycles/Sec.*	Wavelengths. *Metres.*
Long Waves . . .	50–15	6,000–20,000
Medium Waves . .	150–50	2,000–6,000
Intermediate . . .	2,000–150	150–2,000
Short Waves . . .	30,000–2,000	10–150
Ultra-Short Waves . .	above 30,000	below 10

Of these only the long waves and short waves between 12 and 80 metres (25,000 and 3,750 kc/s), are useful for long distance services. The intermediate waves are limited in their range and hence are useful for serving a local area only; for this reason they are used for broadcast purposes and are sometimes referred to as "broadcast waves."

Ultra-short waves have only a very limited reliable range, considerably dependent upon the location of transmitter and receiver.

Since both long and short waves provide a possible means of long-distance communication, a summary of their relative merits is of interest.

In the case of long waves, although attenuation is not great, and the wireless path does not introduce distortion, the signal to noise ratio is poor (due to atmospherics), and everything possible must be done at the receiver to improve this; but unfortunately many methods of so doing result in producing poor signal formation if carried too far. On the other hand, the attenuation remains approximately constant over long periods, and thus the signal arriving at the receiver is uniform

the great magnifying effect of the Beam, and boldly commenced the erection of a chain of Beam Stations in this country and the Colonies. His optimism was fully justified as, within the space of four years (the time required to develop and build the first of these stations), the Marconi Beam produced results which greatly surpassed anything that had been obtained by any other means. Other countries were still obtaining unreliable results and using short waves only for auxiliary stations. Thus was fulfilled the dream of an Empire Chain of Wireless Stations, projected so many years before, but never accomplished.

The success of the Beam has had a profound influence on the history of communication. As far as Great Britain is concerned the Beam was an engineering triumph but an economic tragedy. It was a triumph technically because a wireless communication system was suddenly produced which could work direct to the furthest stations on the earth's surface, over routes hitherto unworkable by wireless means. The system was capable of conveying not only telegraphy at the highest possible speeds but also other forms of intelligence requiring a wide frequency-band.

Economically, the Beam was to prove something of a tragedy for Great Britain. When its capabilities were fully realised and adopted by other countries, Britain lost what she had had up till then, a virtual monopoly of handling world communications through her cable network. This she had obtained chiefly because of the geography of the British Commonwealth which required long trans-oceanic communications, and to the fortuitous circumstance that the Empire could provide so many lands and scattered islands dotted over the surface of the Earth on which the necessary repeater stations could be located. The cost of a wireless beam system is only a fraction of that of a cable, and in the passing years countries which hitherto made extensive use of the British cable routes set up direct beam services of their own, and much traffic was thereby diverted. But since the cable system has many advantages for telegraphy and therefore still forms a valuable national asset, a Government-sponsored merger was formed of all overseas cable and wireless telegraph services in order that the two systems might become complementary.

More recently still this Cable and Wireless Company has been nationalised so that all telegraph and telephone services are now Government-owned and controlled.

An essential link in any radio communication is the path between transmitter and receiver. Since its properties are complex, selective and variable, it may be helpful to consider a broad, if arbitrary, division of the radio wave-range into groups as shown by the Table below, in order to show how short and ultra-short waves fit into the general communication picture.

TABLE I

Classification of Radio Waves.

Classification.	Frequencies. *Kilocycles/Sec.*	Wavelengths. *Metres.*
Long Waves	50–15	6,000–20,000
Medium Waves	150–50	2,000–6,000
Intermediate	2,000–150	150–2,000
Short Waves	30,000–2,000	10–150
Ultra-Short Waves	above 30,000	below 10

Of these only the long waves and short waves between 12 and 80 metres (25,000 and 3,750 kc/s), are useful for long distance services. The intermediate waves are limited in their range and hence are useful for serving a local area only; for this reason they are used for broadcast purposes and are sometimes referred to as "broadcast waves."

Ultra-short waves have only a very limited reliable range, considerably dependent upon the location of transmitter and receiver.

Since both long and short waves provide a possible means of long-distance communication, a summary of their relative merits is of interest.

In the case of long waves, although attenuation is not great, and the wireless path does not introduce distortion, the signal to noise ratio is poor (due to atmospherics), and everything possible must be done at the receiver to improve this; but unfortunately many methods of so doing result in producing poor signal formation if carried too far. On the other hand, the attenuation remains approximately constant over long periods, and thus the signal arriving at the receiver is uniform

in strength. Unfortunately the aerial is the weak link in the chain. It is easily possible to produce fairly efficiently the high-frequency currents required, but very difficult to get an aerial circuit to radiate energy even with a costly mast and aerial system; since also the long-wave aerial cannot conveniently be made directive, even the small proportion of power radiated is not used to the best advantage for point-to-point working. The poor aerial efficiency reacts unfavourably on the transmitter, which must be of large power output, and hence it is difficult to control, besides being costly to build and operate. The modulation frequency is also limited by the time constants of the low-resistance aerials and closed oscillatory circuits, which are essential.

If short waves be used, however, the aerial system becomes much more efficient, and as the wavelength is small the whole aerial can be greatly reduced in size and cost, and also it now becomes practicable to design the aerial to concentrate the radiated energy into a beam, thereby making the best use of it. The transmitter can be greatly reduced in power, it is cheaper to construct and easier to operate, and may be keyed conveniently at high speeds, as at these very high frequencies the time constants of the circuits offer no bar to the highest modulation frequencies desired. The signal-noise ratio is very much better, so that services can be worked with field strengths at the receiving station of very low level, and the attenuation is perhaps rather less than for long waves, although this is governed by entirely different laws.

Unfortunately short waves are subject to rapid physical changes, so that the received signal varies over wide limits from moment to moment, and it is this fading phenomenon rather than atmospheric noise which is the controlling factor in economic design. For many classes of service, it is essential to provide the receiver with some form of compensation to correct for this and to deliver an approximately constant strength of signal irrespective of the fluctuations of input. For various reasons, which will be dealt with later, short waves do not form a distortionless connecting link, and in telephony and "picture telegraphy" the distortion may seriously interfere with reproduction.

Extremely marked diurnal and seasonal changes also occur,

necessitating the use of different wavelengths at different times for the same service. In addition, channels working on routes passing near the magnetic poles are subject, during magnetic storm periods, to such severe attenuation as sometimes to put them out of action.

Thus it will be seen that short waves offer the advantage of a cheap and efficient system, and a good signal-noise ratio, but suffer from a varying attenuation and a tendency for distortion.

Since the field strength used can be much lower than with long waves, the amplification needed at the receiver will be much greater and noises due to local electrical machinery and internal noises present new problems in receiver design.

There is no doubt that the chief contribution to early short wave commercial communication was the " beam." The tremendous power gain of the first array systems used overcame the deficiencies of early transmitting and receiving apparatus and of the connecting link, the ether, and enabled telegraph circuits to operate at speeds not exceeded even to-day. Although the passing years have shown changes in " beam " design, first towards simplification with a view to economy of first cost, and more recently back to elaborate groups of smaller " beams," no advance in efficiency can be recorded. On the other hand, transmitting and receiving gear has improved in efficiency and design very considerably, very important developments being the perfecting of constant-frequency sources, and the technique of feeder and aerial circuit design. The power of short wave transmitters has risen very considerably, and the employment of powers of 100 kW and more has brought many problems to the designer.

Since we have to deal specifically with a communication system using electrical wave energy as a medium, it will not be out of place to discuss first the peculiarities of this branch of electrical engineering, and this can best be done by comparison with the general features of a power system, Table II (page 6) showing the principal points at variance.

The power engineer is accustomed to think largely in terms of a steady state condition, the use of D.C. or a single low A.C. frequency, the transmission of large powers necessitating high power efficiencies, and the avoidance of resonance conditions.

TABLE II

Comparison of Power and Telecommunication Engineering

Power	Telecommunications
Periodic Waves.	(1) Continually changing conditions.
D.C. or a single low-frequency A.C.	(2) Spectrum of frequencies.
Wave-shape often unimportant.	(3) Distortion not generally permissible.
High power transfer, necessitating high efficiency.	(4) In general, power efficiency unimportant as low power involved.
Avoidance of resonance.	(5) Resonance effects utilised.

It is true he has to consider transient effects, particularly when designing switch gear, and it is desirable to preserve as nearly as possible a sinusoidal waveform, but these are rather incidental effects. In telecommunications, things are very different and we will discuss the contrasting conditions given in the second column.

(1) It is axiomatic that no intelligence can ever be conveyed by a system in a steady state condition, but only by transient conditions. Thus the operator who holds down his telegraph key all day is no more transmitting than he who leaves it alone, since no signalling is done until a change occurs. Similarly the transmission along a line of a steady current is not affecting any communication until some change occurs, either by cutting the waves up into some code, or otherwise changing its steady (or in the case of A.C., its periodic) character. This means the communication circuit as a whole has no longer to deal with a periodic wave but with ever-changing, transient conditions. It will be seen, however, in the next chapter that although the analysis of any transient wave-shape is very different from that of a periodic series of similar waves, fortunately the total spectrum width is much the same in both cases, and in consequence telecommunication apparatus can be designed on the basis of steady-state conditions without much loss of fidelity.

(2) Since even the simplest form of signal requires a change of condition, every signal occupies a certain frequency spectrum

which may be very small or large, and in Table III is shown the frequency spectrum necessary for the transmission of the more important types of intelligence.

TABLE III

Frequency Band-width for Different Signals

Type of Intelligence.	Band-width c/s.
Morse or Five-unit Code (to high speed)	0–200
Commercial Telephony	300–3,400
Broadcasting, general	30–5,000
Broadcasting, high fidelity	30–15,000
Facsimile	0–3,000
Television	0–2,500,000

(3) The system as a whole has to be designed to pass the appropriate band of signal frequencies whilst preserving the relative frequency-amplitude-phase composition of the original wave within certain limits, the tolerance allowed being determined by the form of intelligence conveyed and the type of service involved. In addition the output level must be maintained at the correct relative level in spite of probable rapid changes in the intervening medium ; and further, there should be a linear relationship between the output of the system to that of the input at different amplitudes of input. Three of the most serious forms of distortion are :—

(a) Attenuation distortion, due to variation of loss or gain with frequency. This leads to different frequency components of the signal being treated differently, so that their relative amplitudes at the receiving end are not the same as initially produced. This type of distortion must be kept small for all types of signals, except telegraphy.

(b) Phase distortion, due to variation of the propagation time with frequency, causing the relative phases of the component frequencies to change. The effect will be to alter the wave-shape of a complex signal, and is therefore a serious type of distortion in the case of visual signals.

(c) Non-linear distortion, due to transmission system having non-linear characteristics. This gives rise to

the introduction of new frequencies not present in the original transmission. The new frequencies introduced will be harmonics of the original component frequencies, and sum and difference frequencies. Such distortion is serious at all times, particularly so when a circuit is handling two or more signals, as non-linearity in this case produces intermodulation frequencies which cause interference.

Apart from the above, distortion can be caused by the introduction of noise, usually of random character. Quite clearly a strong signal is of little value if there are also unwanted disturbances which produce an output comparable with the signal. On the other hand a very weak signal having a high ratio of signal/noise can be made good use of, if sufficient amplification can be provided. Since random noise is virtually a background covering the frequency spectrum of any type of signal, it cannot be separated from the signal unless its level is much less, and hence the ratio of signal to noise is a most important one and should generally be of the order of 10/1 power ratio (10 db) the actual ratio being dependent upon the class of service.

(4) The power level in a wireless communication circuit varies tremendously from point to point as we pass through successive stages of amplification and attenuation but it will be very small, except at the transmitter, where many kW may be involved. In fact, we are not much concerned with the actual amount of power we have at the different points, except to be sure that the signal/noise ratio is favourable. This means that, except in the final stages of the transmitter, communication circuits are designed to operate at maximum output conditions rather than with high efficiency. Further, there may be very large variations of signal intensity to allow for. Thus, if the system is dealing with sound, for example, these intensity changes may be as great as $10^{12}/1$ (energy), if they are to cover the ear's response at the optimum frequency. Because of these facts, it has been found convenient to adopt, as a transmission unit of gain and loss, one which is exponential (so as to cater for large variations of level) and relative, since we are often not concerned with actual values. The reader will be familiar with the two transmission units in general use :—

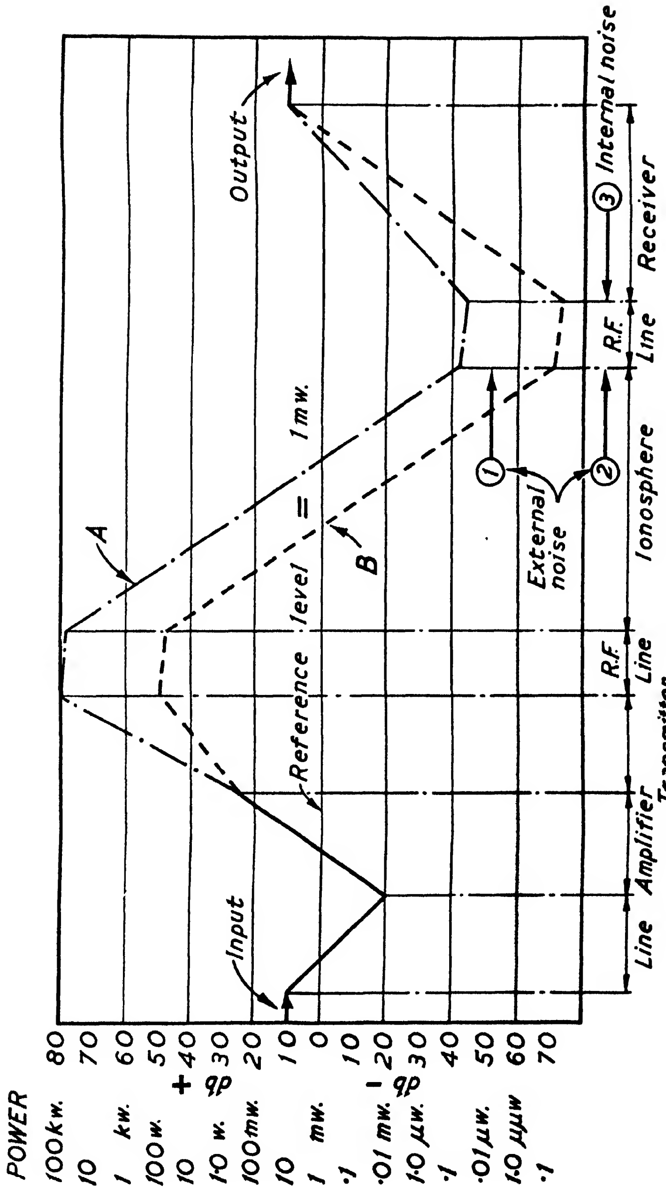


FIG. 1. Radio-communication Power-level Diagram.

$$\text{The Decibel (db)} = 10 \log_{10} \frac{W_g}{W_s}$$

where W_g and W_s are greater and smaller powers,

$$\text{and The Decineper (dn)} = 10 \log_e \frac{E_g}{E_s}$$

where E_g and E_s are greater and smaller voltages.

The telephone engineer is able to think in such terms throughout his system since at no point is any appreciable power involved. But in adopting the transmission unit for a radio system, although the significance of the unit is equally important, we have at the same time to bear in mind that at the transmitter and its associated equipment we may be handling large powers. In such a case we have therefore to consider the power efficiency involved rather than its db loss or gain.

(5) The wireless engineer utilises the phenomenon of resonance very considerably, to provide selective and efficient circuits.

We can illustrate many of the points discussed by considering the level diagram of Fig. 1, for an imaginary wireless telephone circuit. Thus, taking a power reference level of 1 mW, powers above and below this are scaled linearly in multiples of 10, the total power range being from 100 kW to 1/10 $\mu\mu$W., *i.e.*, $10^{15}/1$. Beside the power figures are given the db values above and below the reference of 1 mW. On the assumption that we require a speech output to line at the same level as the input, *i.e.*, + 10 db, the amount of amplification at the receiver will be determined by the level received at its input. If we assume an ionosphere attenuation of 120 db between transmitting and receiving aerials, given by a line of slope as *A* or *B*, then clearly from an economic point of view the lower this line the greater the overall efficiency. For if we could drop the level here from position *A* to *B*, we should reduce transmission power from 100 kW to 100 watts, the received power falling from ·1 μW. to ·0001 of a microwatt, requiring an increased gain at the receiver of 30 db, but at a low level. The factors which determine the limiting condition for reduction of power are external noise-level, H.F. fading, and receiver-circuit internal noise. Thus, if we had an external noise level as shown by (1)

Fig. 1, a ratio of 10/1 signal noise would fix the required input and output as shown by curve A. Whereas, if the external noise was much lower, say as (2), and we had a receiver noise level as given by (3), this would allow the level to be dropped as shown by curve B, assuming there was no variation of level in the ionosphere. Should, however, the wave in the ionosphere be fluctuating 20 db, we would need to raise the level by this amount so that at the minimum of the fading condition we still have a favourable signal/noise ratio.

We can measure set noise accurately, and although external noise varies tremendously with place, time and season, average values for atmospheric have now been charted for most of the earth's surface, and these, together with similar charts for wave attenuation, enable figures to be obtained for an economic basis of design.

To illustrate the importance of keeping in mind the actual power involved when dealing with a level diagram in a radio circuit, let us consider the R.F. lines shown between transmitter and aerial and between the receiver and its aerial. The attenuation of such lines is usually expressed in db but whereas considerable loss is permissible at the receiving end this is not the case at the transmitter. Thus a 2 db loss would be ignored by the receiving engineer, but 2 db at a level of 100 kW, as shown in case A, Fig. 1, means 37 kW, which is much more than could be tolerated by the transmitting engineer.

Reference

TREMELLEN AND COX. "The Influence of Wave-propagation on the Planning of Short-Wave Communication." *J.I.E.E.*, March-April, 1947.

CHAPTER II

SOUND AND VISION SIGNALS

TELECOMMUNICATION involves the transmission of intelligence, the word in the present context being defined as the means by which human beings convey ideas, one to the other. At one time the radio engineer was only concerned with the transmission of the simplest form of signal involving mark and space conditions, but nowadays almost any form of sound or vision signal can be conveyed, and a short discussion on the principal types will therefore be desirable. All forms of intelligence must eventually operate through one or more of our five physical senses, of which only two are of any real interest, sight and hearing. Through our sense of sight we signal by means of signs and images of many kinds; and because of our hearing we have speech and music, and the many other forms of aural signalling. It is difficult to say which to-day is the more important, sight, giving the written word and a heritage of literature together with art in various forms; or hearing, giving us the spoken word with its wealth of emotional quality and its ability to sway the feeling of multitudes.

An objective analysis of either a sound or vision signal is incomplete because the final judgment is a matter for the ear or eye and brain, and is therefore a matter which concerns the individual's characteristics. Thus much of our perception of sound is qualified by our sense of musical appreciation, and our analysis of visible effects may be coloured by our artistic imagination. It is necessary, therefore, in classifying sounds and visible effects, to consider two parallel sets of measurements, one objective, the other subjective.

Objective and Subjective Measurements

The properties of a sound signal, for instance, can be assessed in two ways. The sound may be defined as a particle displace-

ment propagated in an elastic medium, usually air, and is from this point of view specified if we measure the amplitude and waveform of the pressure variation in the medium. Such a measurement, dependent upon instruments measuring physical quantities, is an objective measurement.

From another point of view, however, sound is a sensation conveyed to the brain; its qualities can only be assessed by using the ear and each person's assessment may be different. Such measurements are subjective.

It is a matter of common experience that our senses are unable to measure sensations but can only compare them. This is no doubt due to the tremendous range and the very great accommodating power of the senses to different conditions. Our ideas about the brightness of a room, for instance, depend very largely upon whether we have come from a lighter or darker room. All subjective tests must, therefore, be comparisons carried out under carefully controlled conditions.

Sound

We have already defined sound from the objective and subjective points of view. If we take the acoustic output from a pure tone source and vary it from 0 to 20,000 c/s, the average individual will only experience the sensation of sound (under the most favourable circumstances) between frequency limits of 16 to 16,000 c/s. Between these limits, sounds are divided roughly into two main groups, musical and non-musical, the latter being those to which no definite pitch can be assigned. The random background we call "noise" in a communication channel comes into the latter category, the term noise acoustically however being any unwanted sound, musical or non-musical.

A musical sound has three attributes by which it is recognised, set forth both objectively and subjectively in the table on page 14.

We propose to discuss the above qualities, assuming the reader to have a general knowledge of the functioning of the ear physiologically. Consider a pure tone source connected to a telephone receiver output, whose frequency can be varied and whose acoustical intensity can be controlled within large limits by means of a calibrated attenuator, as indicated in

TABLE IV

Characteristics of a Musical Sound

Objective.	Unit.	Subjective.	Unit.
(1) Fundamental frequency.	Cycles per sec.	Pitch.	Octave.
(2) Intensity or power.	Decibel (referred to 1 mW datum).	Loudness.	Phon.
(3) Waveform.	Harmonic series.	Timbre.	Compared with pure tone.

Fig. 2. Once the apparatus is calibrated against an acoustical reference we can determine the output level electrically in db. from any convenient datum, usually 1 mW. The output so

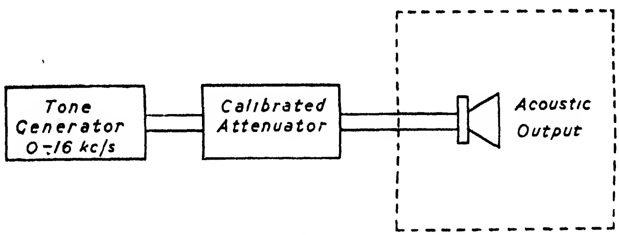


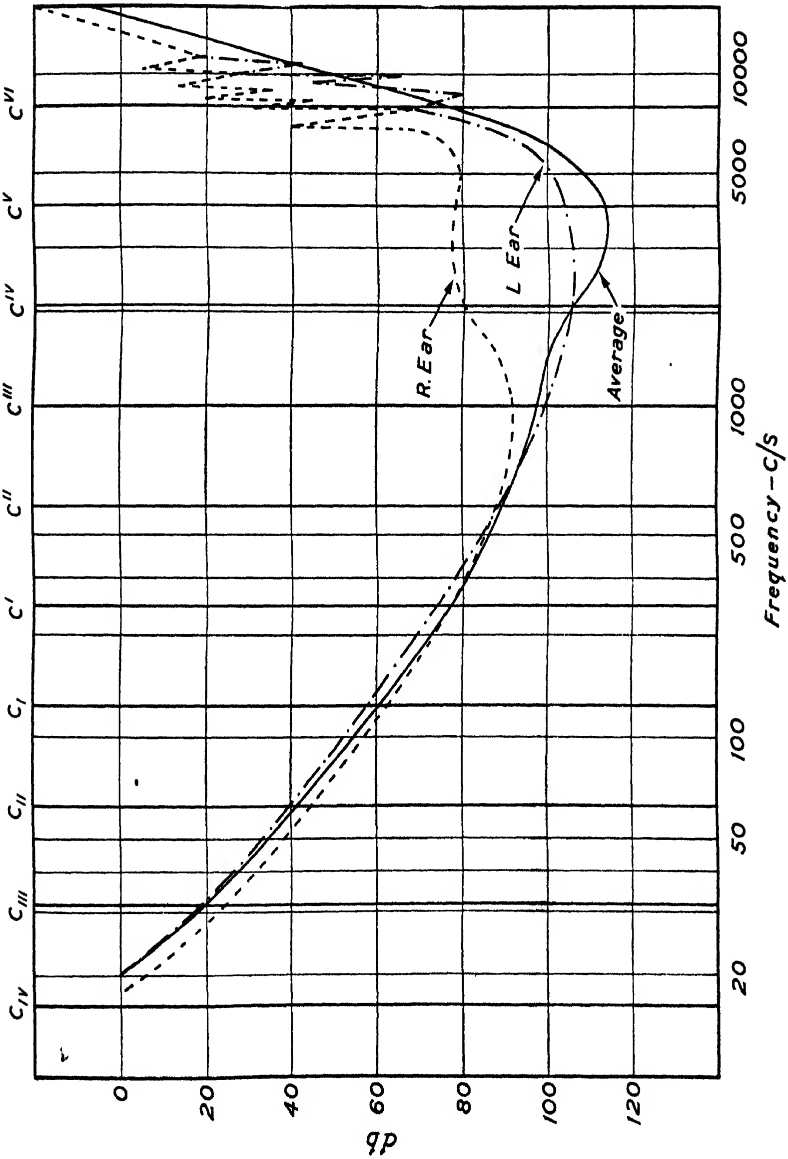
FIG. 2. Acoustical Measuring Apparatus.

produced is then conveyed to the individual undergoing test, from whom the subjective values are determined.

Fig. 3 shows the results of an aural test with each ear of an individual at threshold (dotted curves) and a curve (full line) averaging the results from a large number of individuals. The departure of the dotted curve, Fig. 3, in the case of the right ear shows evidence of a slight deafness in this particular case.

Fig. 4 shows average curves for pure tone sounds at and above threshold, up to a saturation limit of intensity determined by feeling other than sound, this saturation being termed the threshold of feeling.

These curves show the limits of sound perception to vary considerably with intensity, only reaching the maximum range of frequency previously mentioned at a high intensity level.


C_{IV}
C_{III}
C_{II}
C_{I}
C^{I}
C^{II}
C^{III}
C^{IV}
C^{V}
C^{VI}
0
20
40
60
80
100
120
db
20
50
100
500
1000
5000
10000
Frequency—C/s
R. Ear
L Ear
Average

This means the efficiency of the ear at different frequencies is dependent very greatly upon the sound intensity and we

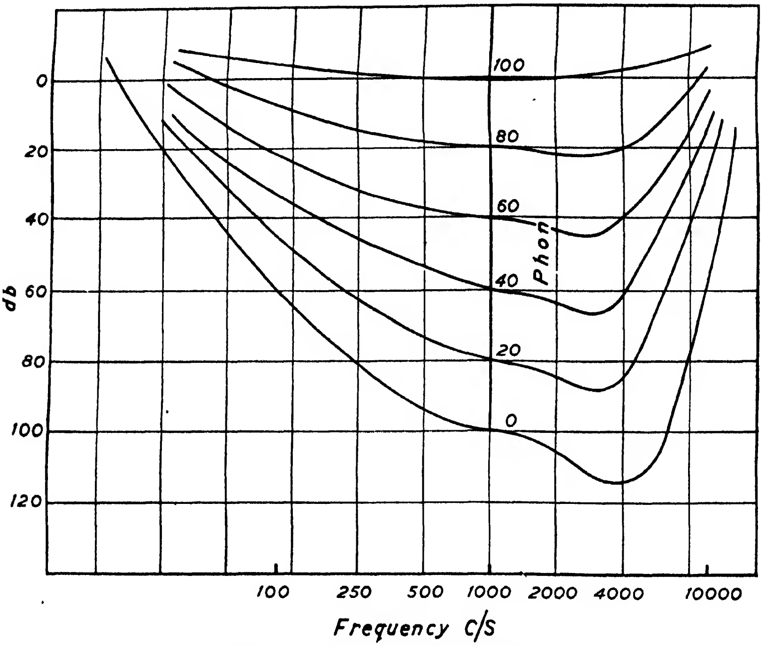


FIG. 4. Aural Loudness Curves.

propose to discuss briefly the ear's reactions to both pitch and intensity changes.

Frequency and Pitch

The frequency of a sound is its measured periodicity, usually in c/s. Such an objective measurement means nothing to the average individual who thinks in terms of the musical pitch of a sound, which is defined as that place on the musical scale to which he can assign it. Some of us have such a small musical sense that we can merely say a sound has a high, low, or medium pitch, whereas others can place any sound fairly accurately. Most people, however, have a similar sense of tonal value to the extent of appreciating the octave interval and its subdivision into the usual musical intervals. And further, if given a reference tone, of assessing pretty accurately

octave intervals above and below, and semi-tone intervals within the octave. These octave intervals, however, appear to the ear to form equal changes, as does a semi-tone anywhere on the musical scale.

The linear scale of pitch is shown at the top of Fig. 3, in octaves above and below middle C, and the frequency scale is shown below, from which we can see that by plotting frequency on a log scale we automatically obtain a linear octave scale.

Since each octave is divided into the same number of semi-tones (12), in appreciating a semi-tone near the bottom end of the scale we are judging frequency differences of 1 c/s or so, whereas at the top end of the scale the difference is some hundreds of c/s, it being important to observe that in every case the percentage change is the same. Most people can judge pitch change to some half a semi-tone, *i.e.*, about 5%, if the tones are offered one after the other, and to a much less percentage difference when the tones are given together, forming a beat. But although our discrimination of tonal change is of this accuracy to sounds near one another within the octave, we have very great difficulty in comparing tones which may be an octave or more apart, particularly if one or both tones are complex in character. Thus the ear's assessment of tonal value is logarithmic, and the musician's scale of the octave, with its almost logarithmic subdivisions of semi-tones is, therefore, a unit which fits in with the natural law of hearing as regards pitch changes.

Intensity and Loudness

The *intensity* of a sound is the measured *output*; *loudness*, on the other hand, is the strength of the sensation produced within the ear. In Fig. 4 equal loudness curves at different levels from threshold to saturation are plotted. The threshold curve is clearly an equal loudness curve, since at each frequency point the intensity has been adjusted to that value where the ear can just sense each tone in question. Equal loudness curves at higher levels are obtained by comparison of the various tones with one of 1,000 c/s raised from threshold in convenient steps by a given number of phons, the phon unit being the same as a db intensity change at this frequency.

Thus to obtain the loudness curve for 20 phons, the 1,000 c/s tone is raised 20 db from threshold. Other tones are then matched by ear, one by one, to appear equally loud to the 1,000 c/s tone, the equivalent attenuation required to achieve this determining the phon value in each case. It is no easy matter to compare the loudness of two sounds of different pitch, particularly at high levels where the subjective tones introduced change their character, and the points obtained are not usually very consistent. Thus the curves shown in Fig. 4 must be interpreted as having a very wide tolerance. They do show, however, that only at a high level of sound *intensity does the ear respond to the whole gamut of frequencies previously mentioned and then with a fairly uniform loudness*; and that as the intensity level is reduced the ear's frequency range becomes restricted more and more and loudness changes become more marked. Observe that the ear's most sensitive region is near 3,000 c/s, that is nearly four octaves above middle C on the musical scale.

Waveform

Sounds may be simple harmonic, in which case the waveform is as shown in Fig. 5, or complex as examples Figs. 6 to 11. But with certain qualifications which are mentioned later,

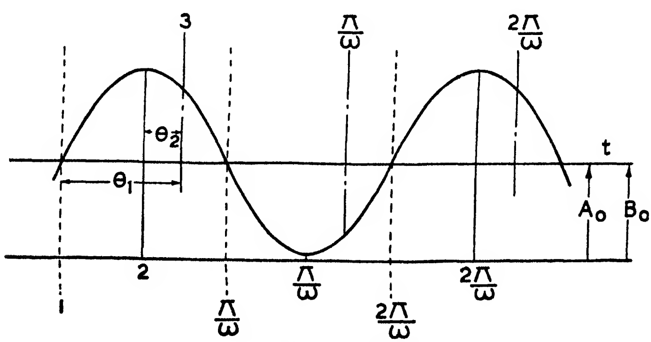


FIG. 5. Harmonic Waveform.

most musical sounds, however complex, usually comprise a central sustained portion with transient beginning and end. We will therefore discuss some features of periodic waves and transient pulses.

A simple harmonic wave may be expressed in various ways depending upon the time axis chosen. Thus in Fig. 5 :—

From a time (1) $F(t) = A_0 + A_1 \sin \omega t$
,, ,, (2) $F(t) = B_0 + B_1 \cos \omega t$

where A_0 and B_0 determine any D.C. term involved should the wave be not purely alternating.

It is important to note that the sine wave is skew-symmetrical or inverted about a vertical time axis of 0, π, 2π, etc., although the inversion is in opposite direction at π, etc. Whereas the cosine wave is symmetrical about a time axis of 0, π, 2π, etc., although the amplitude changes sense at π, etc. From a time as (3), the same wave is neither symmetrical or inverted. It can of course be represented by either a sine or cosine with an appropriate phase factor, but it may also be expressed in terms of the sum of a sine and a cosine, as below, the coefficients having, of course, different values to those above and usually differing from one another.

From a time (3) $F(t) = A_1 \sin \omega t + B_0 + B_1 \cos \omega t$.

If the wave is periodic but complex, it was shown by Fourier that it could be resolved into a series of sine and cosine harmonic terms as given by equation (1), or its alternative form (2) below :—

$$F(t) = A_1 \sin \omega t + A_2 \sin 2\omega t + A_3 \sin 3\omega t - - - - - + B_0 + B_1 \cos \omega t + B_2 \cos 2\omega t + B_3 \cos 3\omega t \quad . \; . \; . \quad (1)$$

$$F(t) = A_0 + A_1 \sin (\omega t + \phi_1) + A_2 \sin (\omega t + \phi_2) + A_3 \sin (\omega t + \phi_3) \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad (2)$$

Equation (2) is of course merely a trigonometrical variation of (1). If such a wave shows symmetry about a time axis, it can then be expressed wholly in cosine terms, and if it shows inversion, by sine terms. We do not propose to discuss the methods of analysis (all of which are tedious) as these are given in a number of mathematical and engineering text-books, but will confine our remarks to the physical interpretation from the communication standpoint.

Fig. 6 shows examples of very complex waves, 6a from a bowed violin and 6b a vowel sound, with their frequency-

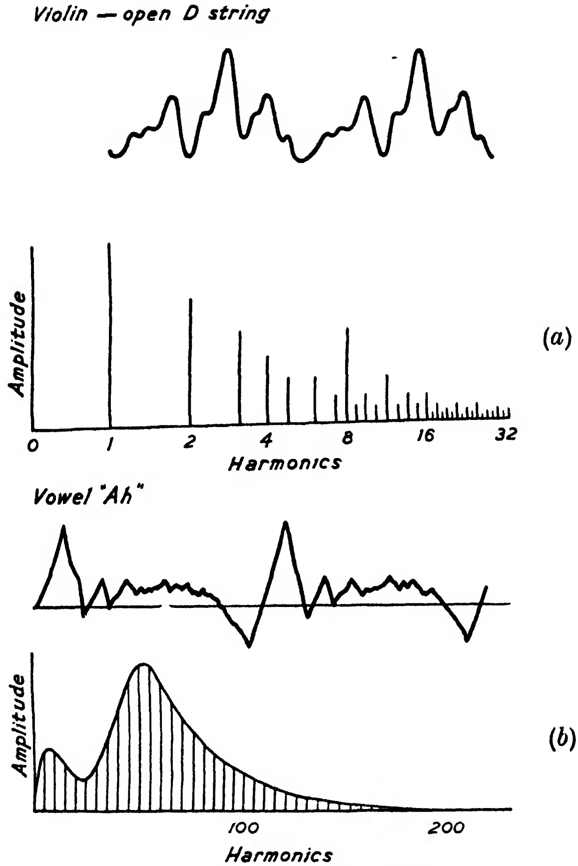
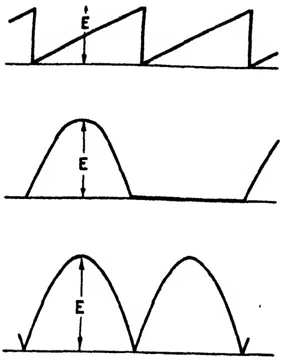


Fig. 6. Complex Sound Waves.

amplitude spectra. With waves of such complexity the coefficients follow no ordered sequence. Figs. 7 and 8 show types of less complex waves very commonly met with and which can be expressed simply in terms of an infinite but ordered series, as shown by the equations given in each case. For convenience, these waves are shown as voltage waves varying with time, but the original Fourier series equations hold, of course, for any periodic quantity and it need not be varying with time.

It should be pointed out that the waveshape is very



$$e = E\left[\frac{1}{2} + \frac{2}{\pi}\left\{\sin \omega t - \frac{1}{2}\sin 2\,\omega t + \frac{1}{3}\sin 3\,\omega t - \frac{1}{4}\sin 4\,\omega t + \cdots\right\}\right]$$

$$e = E\left[\frac{1}{\pi} + \frac{1}{2}\sin \omega t - \frac{2}{\pi}\cdot\frac{1}{3}\cos 2\,\omega t - \frac{2}{\pi}\cdot\frac{1}{15}\cos 4\,\omega t - \frac{2}{\pi}\cdot\frac{1}{35}\cos 6\,\omega t - \cdots\right]$$

$$e = E\left[\frac{2}{\pi} - \frac{4}{\pi}\cdot\frac{1}{3}\cos 2\,\omega t - \frac{4}{\pi}\,\frac{1}{15}\cos 4\,\omega t - \frac{4}{\pi}\cdot\frac{1}{35}\cos 6\,\omega t \cdots\cdots\right]$$

FIG. 7. Complex Wave shapes.

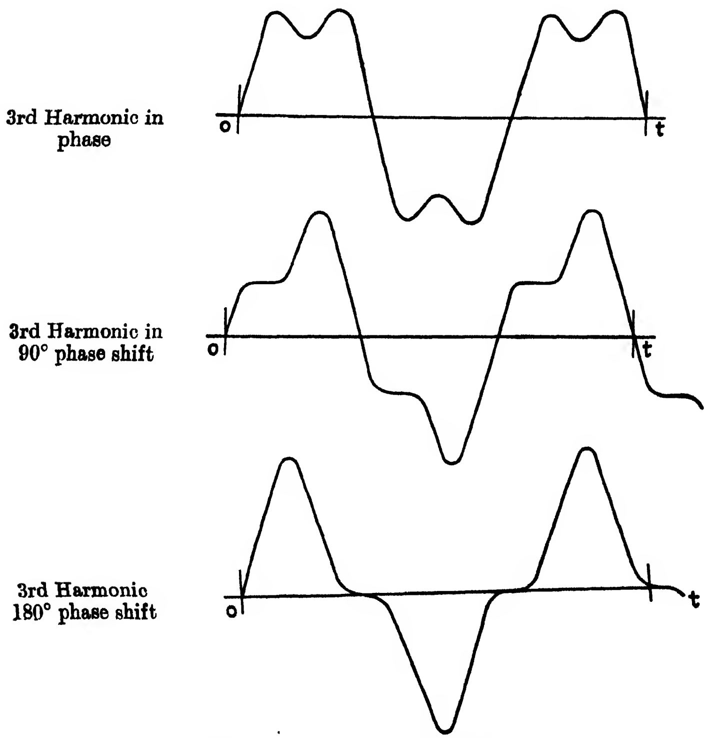
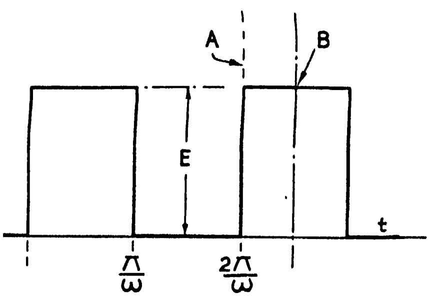


FIG. 8. Effect of Varying Phase of Third Harmonic.

dependent upon the phase relationship of the component frequencies and very great changes can be brought about by small phase shifts. This makes it difficult for even an experienced observer to know by inspection what harmonics are present, except in very simple waveforms. Thus with a fundamental and third harmonic alone, the envelope changes through a series, from which three shapes only have been selected as in Fig. 8. It is found that the ear is incapable of recording phase relationships in musical sounds and is only concerned with the frequency-amplitude composition. Thus phase distortion is of small account, having an influence only with sounds of a percussion type.

All tho waves shown exhibit similarity in that to produce



Time measured from A

$$e = E\left[\frac{1}{2} + \frac{2}{\pi}\left\{\sin \omega t + \frac{1}{3}\sin 3\,\omega t + \frac{1}{5}\sin 5\,\omega t \cdots\cdots\right\}\right]$$

Time measured from B

$$c = E\left[\frac{1}{2} + \frac{2}{\pi}\left\{\cos \omega t - \frac{1}{3}\cos 3\,\omega t + \frac{1}{5}\cos 5\,\omega t \cdots\cdots\right\}\right]$$

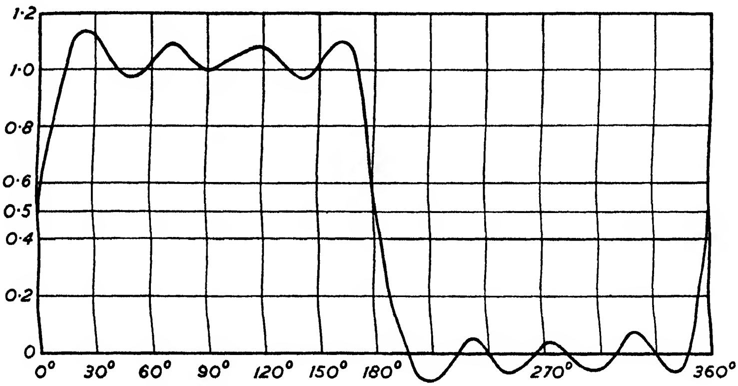
FIG. 9. Square Wave.

them at a given fundamental frequency necessitates a frequency spectrum very many times greater. Thus with the wave of Fig. 6, the band-width is some 30 to 40 times the fundamental frequency.

With waves of simpler type, however, examples of which are given in Figs. 7 and 9, although they are expressed mathematically only with an infinite series, because the harmonic amplitudes are inversely proportional to frequency, the coefficients quickly diminish to negligible values. In consequence, such waves can be produced approximately, using only a restricted frequency band. Thus with the square wave of Fig. 9, although an infinite series of odd harmonics are required, a fair wave-shape will be obtained by a series restricted to the seventh, as can be seen by comparing Fig. 9 with Fig. 10.

This square wave, Figs. 9 and 12a, and its change to a series

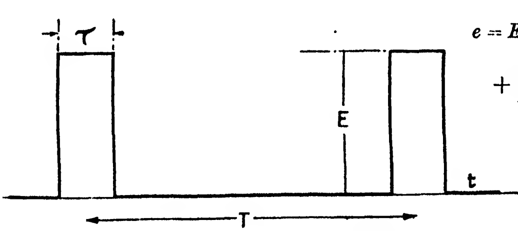
of rectangular pulses, Figs. 11 and 12b, c, d, is of considerable importance in many branches of communication and we propose to discuss it at greater length. Since such a wave shows



$$e = E\left[\frac{1}{2} + \frac{2}{\pi}\left\{\sin \omega t + \frac{1}{3}\sin 3\,\omega t + \frac{1}{5}\sin 5\,\omega t + \frac{1}{7}\sin 7\,\omega t\right.\right]$$

Fig. 10. Square-wave; Fundamental and Harmonics up to Seventh.

inversion about a time axis of A, and symmetry at B, it can be expressed either wholly in sine or cosine terms as given by the two expressions in Fig. 9. The amplitudes are the same in each case except that those of the cosine series alternate in sign.



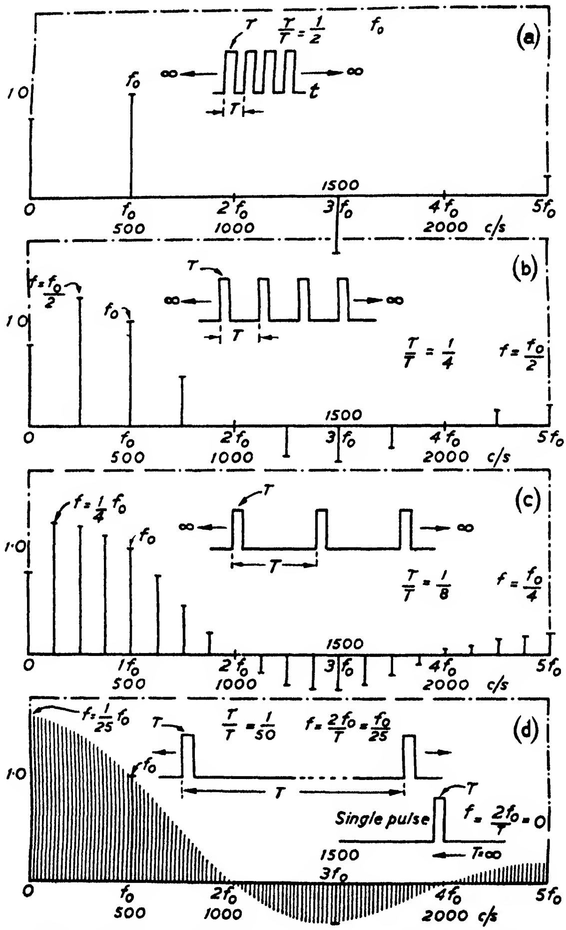
$$e = E\left[\frac{\tau}{T} + \frac{2}{\pi}\left\{\sin\frac{\pi\tau}{T}\cos\frac{2\pi}{T}t + \frac{1}{2}\sin\frac{2\pi\tau}{T}\cos\frac{4\pi}{T}t + \frac{1}{3}\sin\frac{3\pi\tau}{T}\cos\frac{6\pi}{T}t + \cdots\right\}\right]$$

Fig. 11. Rectangular pulse.

Thus if we have a square wave of frequency $f_0 = 500$ c/s, say, the frequency-amplitude coefficients for the cosine series will be as shown in Fig. 12a, for the first three terms. This indicates that there are discrete energy concentrations at the fundamental f_0 and the odd harmonics, but none at the even, or at inharmonic frequencies. There is a D.C. term since we are dealing with a pulse, and in Fig. 12a we have denoted the pulse duration

time as τ and the pulse repetition time T, which in this case is, of course, twice τ.

Let the repetition time T now be increased, leaving the



Contour is spectrum for single pulse of peak duration τ

FIG. 12. Effect of Varying Recurrence Frequency.

duration time τ the same. This means the repetition frequency f will become less than f_0, and the wave changes from square to rectangular as shown by Fig. 11 or 12b, c and d. The expression for such a rectangular wave will now be in terms of the

repetition frequency f, not f_0, and is given in general form in Fig. 11, applicable to any ratio of the times τ/T. Observe that when $T = 2\tau$, $f = f_0$, and this expression reverts to that given for the square wave of Fig. 9.

Using this expression and taking examples of $f = \frac{1}{2}f_0$, $f = \frac{1}{4}f_0$ and $f = 1/25f_0$, the coefficients for each case are as shown in Table V, and cover a frequency band up to the third harmonic of the original square wave of frequency $f_0 = 500$ c/s. It will be seen from this table that as the pulse duration time is increased progressively, since energy concentrations appear at the harmonic frequencies of f, the greater T, the closer these harmonic frequencies become and the smaller the actual amplitudes of each coefficient ; but the envelope shape obtained by joining up the peaks of the coefficients in each case conforms to the framework of the original square wave series of frequency f_0, determined by twice the pulse duration time τ. In order to show this point we have increased the value of the coefficient in each group relatively so that in each case (Fig. 12), the frequency coefficient at f_0 is always unity. These envelopes it will be observed still pass through zero at frequencies of $2f_0$, $4f_0$, etc., and approximate to maxima at the odd harmonics of f_0, but that as T is lengthened, the spectrum extends towards zero.

It will be observed from Fig. 12 that when $T \gg \tau$ the harmonics of f are very nearly equal for the first few terms, indeed there is no great variation from f up to f_0, indicating a considerable proportion of the total energy in this band. And that although energy concentrations are now only harmonically related to f and not f_0, nevertheless we require the extension of the series to harmonics of the duration frequency f_0 in order to form the pulse shape. For it is the duration time which determines the high frequency band required, the repetition time determining the extension of this band towards the lower frequencies. An example will make this point clear. Thus if a 1 millisec pulse is required $f_0 = 500$ c/s, and if the shape fidelity to the 7th harmonic is sufficient, frequencies up to 3,500 c/s will be necessary. If such pulses are produced 100 c/s, the band will extend downwards to 100 c/s and we need harmonics up to the 35th. Whereas if pulses are only produced 10 times a second, the band extends down to 10 c/s and har-

TABLE V

Analysis of Pulses

Harmonics of f_0	$\frac{\tau}{T} = \frac{1}{2}$						1				2				3				4				5
Coefficients .							1·0				0				·33				0				·20
Harmonics for $f = \frac{1}{2} f_0$.	$\frac{\tau}{T} = \frac{1}{4}$				1		2		3		4		5		6		7		8		9		10
Coefficients .					·707		·50		·236		0		·141		·166		·101		0		·078		·10
Relative coefficients × 2 .					1·414		1·0		·472		0		·283		·33		·202		0		·156		·20
Harmonics for $f = \frac{1}{4} f_0$.	$\frac{\tau}{T} = \frac{1}{8}$			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Coefficients .				·375	·354	·308	·250	·186	·118	·054	0	·043	·071	·084	·084	·072	·051	·026	0	·023	·040	·049	·050
Relative coefficients × 4 .				1·496	1·412	1·23	1·0	·74	·47	·212	0	·170	·283	·33	·33	·286	·202	·100	0	·08	·156	·196	·200
Selected harmonics for $f = \frac{1}{25} f_0$.	$\frac{\tau}{T} = \frac{1}{50}$	1	6	10	15	20	25	30	35	40	50	55	60	70	75	80	85	90	100	110	115	120	125
Coefficients .		·063	·062	·058	·054	·047	·04	032	·023	·015	0	·006	·009	·013	·013	·012	·009	·007	0	·005	·007	·0079	·008
Relative coefficients × 25 .		1·57	1·52	1·45	1·35	1·19	1·0	·792	·577	·372	0	·141	·245	·33	·33	·297	·238	·164	0	·134	·175	·198	·20

monics of 10 c/s up to the 350th will have to be passed to obtain the same fidelity of wave-shape.

Transients

A consideration of the above enables us to anticipate what will happen to the frequency spectrum if we make the repetition time infinitely long, so as to produce a single pulse. The frequency spectrum now extends from zero to infinity, but each component is infinitely small in amplitude but infinitely close together. If, however, the components within any band are added, it will be found that the distribution of amplitudes over the frequency scale still follows the contour previously shown in Fig. 12d.

It can be shown that the expression for such an isolated pulse is an integral known as the Fourier Integral, which in terms of τ and T becomes :—

$$F(t) = \frac{2E}{\pi}\int_0^{\infty}\frac{\sin\omega\tau}{\omega}\cos\omega\left(t-\frac{\tau}{2}\right)d\omega.$$

It is simply the integration of the coefficients from zero to infinity for a wave pulse of a certain duration time τ and a zero repetition frequency f.

Single transients of other shapes would reveal a similar characteristic, namely a continuous energy/frequency spectrum from zero to infinity, having a contour which would link up with the discrete frequency points determined by the Fourier series analysis for that particular wave-shape treated as one of a series as f frequency f_0 as in previous example.

Although the spectrum is now infinite, the energy level at high frequencies is very low. The important parts are contained : first over a spectrum from zero to f_0 whose coefficients are more or less uniform in amplitude, necessary for building up a single wave, or the series if the pulse is one repeated at long intervals. And secondly, the continuous spectrum from f_0 upwards whose coefficients form the contour passing through the discrete frequency points of the series analysis, this second part building up the shape of the pulse.

Between the extremes of a single wave and the series we have the case of a group of waves, few or many in number. The

spectrum of any particular case can be found by appropriate modification of the coefficients shown in the above integral which have to be multiplied by a factor which is dependent upon the number of waves in the group. The resulting coefficients for a group of nine waves of pulse duration τ and pulse repetition $T = 2\tau$ are as shown in Fig. 13, where the original discrete frequency f_0 coefficients are shown and the continuous contour. It will be seen that the energy is concentrating more towards the original discrete frequency points, and if the number of waves in the group is increased still more, the inharmonic frequencies are still further reduced so that the

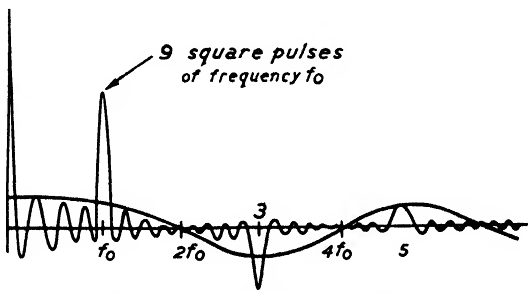


FIG. 13. Limited Pulse Train.

spectrum is now almost approximating to that of a continuous series.

In communication work we are handling an ever-changing waveform and the type of transmission therefore involves continuous rather than discrete frequency spectra. The transient itself has no character of its own but merely changes the discrete nature of an ordered frequency spectrum to a continuous one covering much the same frequency band overall. As an example, the frequency spectra of most vowel sounds lie below a maximum value of some 3,000 c/s. It is found that with a system having a sharp cut-off at this frequency, provided there is no low frequency cut-off, the transients of speech represented by the many consonants are still transmitted without loss. Thus it is of interest to point out that acoustical apparatus designed on steady-state assumptions is often capable of dealing with most of the transient conditions that arise, a somewhat fortuitous circumstance.

Timbre

Subjectively we distinguish between sounds of similar pitch and loudness by what is known as timbre, that is the character of a sound compared with that of a pure tone. There are two separate qualities which have a major influence on the character of a musical sound. First, the richness of its harmonic content, which can make either a consonant or dissonant sound; second, the relative duration of the central sustained portion of the sound to its transient beginning and ending, for convenience termed " attack " and " decay."

Dealing first with consonant sounds (agreement of tones) and dissonant sounds (harshness of tones), both are equally important. Thus we have instruments whose sounds are wholly consonant, and others in which the characteristic

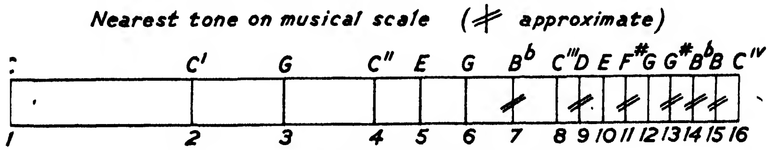


FIG. 14. Musical Scale and Harmonic series.

feature is derived largely by the dissonant element. Both are subjective qualities with which we will deal briefly.

If we consider a complex mechanical oscillating system sustaining a tone, such as a reed pipe, or the human voice sounding a sung vowel, such a system will produce a fundamental tone and a group of harmonic tones. In addition, there may be energy produced at inharmonic frequencies, but at a lower level. Considering the harmonics, these are an arithmetic series, and we have shown them in Fig. 14, set out for a fundamental tone of *C*, below a scale of octaves. From this it is seen that the harmonics 2, 4, 8, and 16, etc., are coincident with the octaves of the fundamental, harmonics 3, 6, 12, etc., coincident with a justly tempered musical fifth, *i.e.*, *G* in the scale chosen, and harmonics 5, 10, 20, etc., are coincident with a justly tempered third, *i.e.*, *E* for the scale chosen. Since *C*, *G* and *E* and their octaves are all consonant, the first six harmonics of the fundamental will be consonant. But the 7th harmonic approximating to B♭ on the chromatic scale will be very dissonant, as will all the odd harmonics after

the 7th, in more or less degree depending on their beating effect with the consonant tones. Multiples of these odd harmonics, *e.g.*, 14th, 18th, etc., will also be dissonant. This means that instruments designed to produce consonant sounds will suppress harmonics after the 6th, the diminishing series after the sixth leading to the introduction of a dissonant element which may merely sharpen the tone, or introduce a definite strident quality so characteristic of reed instruments of clarionet and organ trumpet type. Here the dissonant harmonics are mostly 7th, 9th and 11th.

The subjective character of a sound is determined not only by the number of harmonics present, and their relative amplitude (their phases are immaterial since the ear takes no account

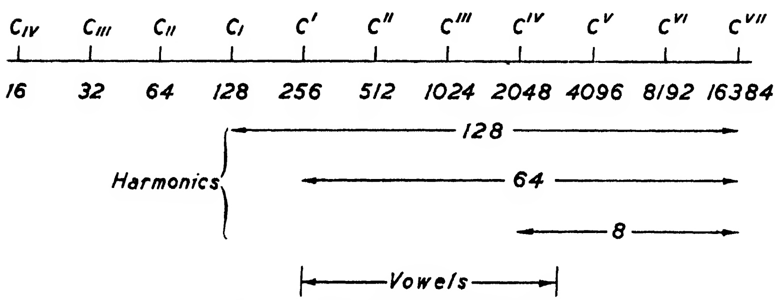


FIG. 15. Aural Harmonic Ranges.

of phase), but also upon the position of the sound series on the musical scale. For instance a low tone rich in harmonics has a very different sound to one of similar character played much higher up the scale. This is due to the ear's performance. Thus in Fig. 15 we show the number of possible harmonics that can be heard from fundamental tones at different points on the musical scale. If, for instance, the fundamental is at middle C, we could hear harmonics up to the 64th, and those up to the 10th would be exaggerated in importance due to the ear's increasing response up to this frequency. Whereas if the fundamental is 3,000 c/s, only four harmonics could be heard and these will be greatly attenuated by the ear's lessening performance.

Thus we find the instruments rich in harmonics have a compass at the lower end of the scale, the top being reserved

for instruments having a very restricted harmonic range, as flutes and piccolos.

The second quality of sound we mentioned is the relation of the sustained portion to its attack and decay time. Thus a string struck as in a piano has an attack determined by the condition of the felt on the hammer, and a decay determined by the damper, this decay being usually long if the damper is left off. If the same frequencies were heard as a sustained sound it would be quite unrecognisable. Thus with plucked and struck strings it is the type of attack and decay which make for recognition rather than the harmonic series. Whereas with most pipes, the sound is nearly wholly sustained with abrupt attack and decay, although there are a few organ stops in which the attack and decay time has some considerable influence. Speech articulatory sounds are a good example. We may consider the vowel as the middle sustained portion of a sound having a defined harmonic series, the consonants at beginning and end determining the type of attack and decay, and having a major influence on the type of articulatory sound as a whole.

Speech

Having discussed musical sounds in general we propose now to discuss briefly one of particular importance, speech, which is of two types, phonated, and unphonated or whispered, speech. The latter is obtained by opening the larynx so wide that the vocal chords do not produce any sound as the air passes through, and thus only articulation from the throat and mouth passages is heard. With phonated speech, the articulatory sounds from the mouth are impressed on a larynx pitch, which thus acts as a carrier, reinforcing the articulation so as to carry it to greater distances besides giving it an emotional quality which is lacking in whispered speech. This larynx pitch is somewhat ill-defined in ordinary speech, but can be controlled to a clearly defined pitch as in singing.

Although normal speech sounds are mostly phonated, the larynx is not really an organ of speech, since it contributes a negligible amount to the intelligence being transmitted, the articulatory sounds proper being produced by the throat and mouth passages which form an acoustical resonating system

giving a sound dependent upon the shaping of the mouth, the position of the tongue, and the forming of the lips and teeth. Since the mouth is capable of assuming a large variety of shapes and can make big changes of volume, the number of articulatory sounds it is possible to form is extremely great.

In general articulatory sounds are divided into two main groups, vowels and consonants, although phonetically these are subdivided into a number of divisions as diphthongs, semi-vowels, plosives, etc. Vowels can be defined, not as alphabetically by the letters a, e, i, o and u (in fact i is not a vowel but a diphthong), but as the open, sustained sounds of speech. Thus they are apart from the closed, sustained sounds, which are the non-musical fricatives or sibilants as " s," " th," " sh," " z," etc., and the vowels come within the category of the central sustained part of a musical sound previously defined, and have, therefore, an approximate harmonic frequency spectrum.

TABLE VI

Frequency Band-widths of Vowels

Vowel Order.	Phonetic Sound.	Example.	Frequency Band.		Inverted Order.	Recognition.
			Lower.	Upper.		
1	oo	who	250–450	2,000–3,000	13	Good
2	ō	to	300–600	2,000–2,500	12	Good
3	oh	know	400–650	1,800–2,500	11	Fair
4	aw	small	500–800	900–1,100	10–13	Poor
5	o	frogs	600–900	1,000–1,200	10–12	Poor
6	ah	pass	700–1,500	700–1,500	8–12	V.poor
7	uh	up	700–900	1,100–1,400	10–12	Poor
8	er	wher	1,000–1,500	1,400–1,900	8	Good
9	a	land	600–800	1,700–1,900	7–9	Poor
10	e	men	450–700	1,800–2,000	5–7	Poor
11	eh	wade	400–600	1,800–2,500	3	Fair
12	i	in	300–600	2,000–2,700	2	Good
13	ee	reeds	250–500	2,000–3,000	1	Good

Notes. Predominating bands underlined.
Pitch of all phonated inverted sounds high.
Fricative and all stop consonants invert badly.

The number of possible vowels that can be spoken or recognised depends upon the acuteness of hearing of the listener

and upon the ability of the speaker to hold his mouth into a number of discrete positions between two extremes. Actually most English-speaking people are able to speak and differentiate between about 13 different vowels as given in Table VI, dialect differences being not so much due to any inability of any community to appreciate the various sounds, but that they use the same vowels in a different context. From the table we see that the extreme vowels are "oo" and "ee," and that phonetically "ah" is the central vowel. This is formed by opening the mouth as wide as possible, keeping the tongue as low as possible, thus forming a single large cavity-resonator, whose frequency-amplitude spectrum is as shown in Fig. 16, the peak amplitude occurring about 900 c/s. From this central vowel position we can close the mouth in one of two ways: by pushing the tongue upwards and back, thus dividing the mouth into two cavities connected by a narrow passage between tongue and palate, at the same time rounding the lips and usually keeping the teeth apart; or thrusting the tongue forwards and upward again dividing the mouth into two cavities, but this time the teeth are closed rather than the lips. In carrying out the first action we pass through the series of vowels from "ah" to

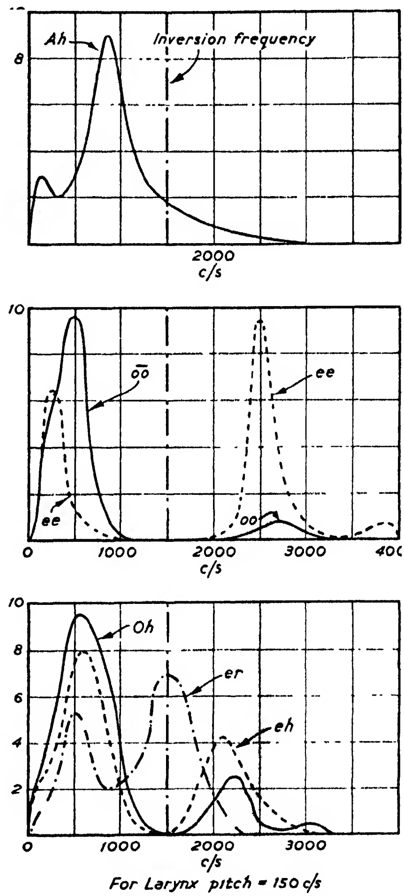


FIG. 16. Spectra of Vowel sounds.

"oo" and from "ah" to "ee" with the second action, the number of sustained pauses we can make determining our ability of vowel production.

Either of these actions result in producing two series cavity-resonators, the first having a low predominating frequency, the second a high one, the double resonance frequency groups being shown in Table VI, with the principal group underlined. The frequency-amplitude spectra for one or two of these vowels is given in Fig. 16, these being phonated with a pitch of 150 c/s, but the shape of the spectra will be much the same whatever the pitch tone used. It will be seen that the upper limit of frequency does not exceed 3,000 c/s in any case, and it is to be remembered that although the intensity level of the higher frequencies is small in some cases, these frequencies coincide with the ear's most sensitive region. Whereas the ear is fairly insensitive at the pitch fundamental of 150 c/s.

Dealing with the consonants, we may regard these as the transient parts of articulation, and thus playing a very large part in determining the character of the sound as a whole. Of these a number are not much removed from a vowel as the "m," "w," "y," "l" and untrilled "r" and have, in fact, a similar spectrum range, but the so-called stop consonants as "b" and "p" have a very steep wavefront and are in the nature of a pulse, thus requiring, as we have seen, an extension of the spectrum to lower frequencies and wide enough to accommodate the series for the pulse shape. In general it is found that the frequency spectrum that accommodates the frequencies of the vowels will also be sufficient for the transient consonants, the latter merely adding some considerable inharmonic energy. The chief trouble is that the energy level of the average consonant is extremely low compared with that of the vowel. It will be seen later, that wireless telephony employs apparatus switched by the voice. Hence, unless the initial consonant has sufficient energy to switch the gear, articulation will be mutilated.

Apart from the larynx pitch, the band-width required for speech articulatory sounds lies in a band from 250 to 3,000 c/s. The addition of the larynx pitch sound does not destroy the mouth resonance frequency bands (assuming the articulation is retained), but adds an extensive frequency spectrum which

is large compared with the articulation energy, and varies greatly with the individual and whether he is speaking or singing. Male speech pitch may be as low as 120 c/s and female of the order of 350 c/s, and the singing compass usually ranges over about two octaves. This indicates that the total frequency spectrum ranges from some 100 to 3,000 c/s, but it has been found that we can restrict this at the lower end by cutting off frequencies below some 250 c/s without destroying quality or articulation efficiency, but we cannot afford to restrict the top end to any extent. It is of interest to note that by cutting the bottom below 250 c/s we are eliminating a very large percentage of the total speech energy caused by larynx sound, but this appears to make not only no difference to the articulation (which is to be expected), but small difference to the apparent voice pitch even with low male voices. This, it has been assumed, is due to the fact that the voice is very rich in harmonics, particularly the male voice, and the subjective difference tones from these give a sufficient pitch content to satisfy the ear. As was mentioned earlier the ear is not anyway very sensitive to octave differences.

Summarising, we can state that the narrowest frequency spectrum that can be tolerated for commercial speech is one from 250 to 2,750 c/s, although extension above this limit is to be desired and the international standard is 300—3,400 c/s.

Inverted Speech

In order to achieve secrecy of speech transmission, a process known as speech inversion is sometimes carried out. This consists of inverting the frequencies within the speech band, thus producing a different set of phonetic sounds which, without some considerable experience, are unrecognisable. The method whereby this is accomplished will be discussed later in the book. The inversion frequency chosen is 1,500 c/s so that any component frequency in a speech sound of 1,500 c/s will be left unchanged, those frequencies below being converted to a corresponding frequency above and *vice versa*, so that the 3,000 c/s point of speech becomes the zero point of the inverted speech. We are thus limited to a band width of 3,000 c/s.

The second part of Table VI shows briefly the results of

phonetic tests on inverted vowels and this together with the curves in Fig. 16 should suffice to indicate why certain vowels invert easily and others are more difficult, the inversion frequency being indicated in Fig. 16.

The first point of interest is that with straight speech not only is the larynx pitch energy concentration well below the inversion frequency point, but also the lower mouth resonance energy. This means that on inversion the main energy of speech will be transferred from concentrations well below the ear's most sensitive frequency point to a region very near it, and the result is that inverted has an exaggerated high pitch. Dealing with the vowels in particular, it is seen from Table VI, that the vowels 1, 2, 3 are inverted forms of 13, 12, 11, and *vice versa*. That they should so invert will be seen from the frequency spectra, which show that the double resonance energy peaks lie roughly equally either side of the inversion point. The vowel " er " remains unchanged, indicating general symmetry about the inversion point of the frequency band. The single resonance vowel " ah " cannot be inverted since its resonance point is about 1,000 c/s and the inversion point lies at 1,500 c/s, and the same applies to those vowels lying near " ah " which have double resonance peaks near the 1,000 c/s point. Thus of the 13 vowels, seven are merely transposed, the other six being turned into sustained sounds which are not capable of enunciation by the ordinary individual, with ordinary speech training. The fact that some of the consonants such as diphthongs and semi-vowels are to a great extent harmonic in character means that some of these also merely transpose upon inversion. Thus " y " and " w," " m " to " n " and the diphthongs " ow " to " ie " and " oi " to " ew." The stop consonants, however, all transpose badly as would be expected, and it is largely on account of these that inverted speech is difficult to recognise and articulate. It should be pointed out, however, that inverted speech is merely an unusual form of phonetical sound, and can be learned by individuals having an acute ear, and with sufficient patience ; and even spoken with a fair degree of success, as is evidenced by the ability of operators familiar with speech inversion circuits. This means that speech inversion is not a secret type of privacy in itself and further modifications are usually made.

Vision

As there is no immediate prospect of the transmission of colour pictures, whether still or moving, we do not propose to discuss chromatic light, except to mention in passing that, just as musical sounds have three attributes, so chromatic light has three qualities of a somewhat similar nature having both objective and subjective values.

We are at present therefore concerned with the transmission of monochromes, that is the equivalent of black and white in the case of print and line drawings and the like, or a series of greys as with half-tones. The subject-matter being transmitted need not necessarily be black and white, nor need the received picture. Thus, if a pick-up device is used which responds to all colour more or less equally, its response to a coloured scene will be graduated in terms of monochromatic light, and at the receiver end such a scene need not be reproduced in terms of black and grey, but may, in fact, appear in any convenient monochrome, as, for instance, the blue or green so often used with the ordinary cathode ray screen.

The transmission of either still pictures (facsimiles) or moving pictures as in television involves a special technique since we cannot reproduce directly the effect of a visual scene as it appears in the retina of the eye. For, with direct vision, the retina of the eye receives simultaneously an infinite mosaic of light and shade, the separate parts of which are simultaneously conveyed to the brain as a complete picture. Thus only if we had an infinite number of separate transmitters each connected to its own receiver, and each single circuit being detailed to pick out a particular spot of the scene, could we reconstruct that scene as it appears to the eye. Although such a system has been proposed it is too crude to merit attention, and all modern methods of facsimile and television adopt a method of scanning the scene by a moving pencil, in such a way that at any instant of time only a single signal has to be passed from transmitter to receiver, the individual signals being synthesised at the receiver in the correctly ordered manner to yield the required visual scene being scanned at the transmitter.

With a still picture the time taken to scan the complete fascimile is immaterial and will depend simply upon circuit or economic considerations. But with television, since there will

usually be movement in the scene, each separate picture must be scanned in a time short compared with the persistence of vision. Actually the problem of flicker and not the persistence of vision, determines the picture speed, as it does in the cinematograph. The equivalent of some 48 complete pictures a second must be scanned as a minimum if flicker is to be reduced to a negligible amount.

There are many ways in which a subject may be lighted and scanned. In facsimile transmission, the normal method is to illuminate the subject only by a single spot of light, reflected light, if any, from the surface under the spot at the time being passed to a photocell connected to the input of an amplifier. The spot of light then traverses the area of the picture in a series of parallel scanning lines so that every part of the subject has been covered. Usually the picture is wrapped around a drum, which spirals past a fixed light spot, or the light source is arranged to rotate within a drum holding the subject, the latter method being preferable as the subject-matter can be set up and removed without stopping the rotating machinery. This actually is an important point as the scanning devices have, of course, to synchronise at transmitter and receiver, and continuously running machinery is more or less essential if exact synchronisation is to be obtained. Theoretically either method is the same, since the area will be effectively scanned, provided that the dimensions of spot are related to the traverse in such a way that the scan line areas are continuous. The amount of detail that will be obtained is clearly inversely proportional to the size of spot, but, since the smaller the spot the greater the number of scan lines, a longer total time will be required for scanning the same area. Most facsimile work is done with pictures up to a maximum dimension of about 20 × 15 cm. and a spot size of the order of 2 mm. is used. The scan speed depends chiefly upon wireless circuit conditions, and as facsimile transmission is carried on over great distances, where conditions are very variable, the speeds used in practice vary greatly. If the subject to be scanned is a black and white, then ordinary D.C. keying methods can be employed, since the light is merely cut off and on, but if gradation of light has to be transmitted, this involves a modulation system. Actually a D.C. system has been used for such subject-matter by the Ranger facsimile

system, the gradation of the subject being obtained by the size of the transmitted marking dot. Thus for black the dots will merge forming a continuous mark, whereas for white there will be only a spacing condition, the gradations between being covered by the change in the length of the mark to space

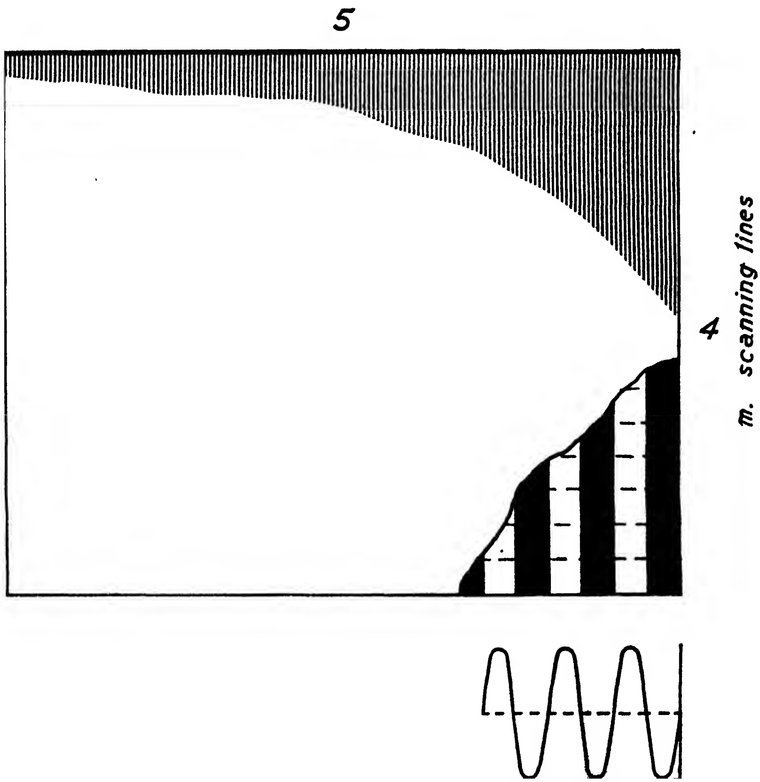


FIG. 17. Illustrating Picture Transmission (corner of picture shown enlarged).

forming smaller and smaller dots as the grade becomes lighter. The effect is equivalent to the half-tone technique employed by printers. Alternative methods utilise light which is varying in intensity at a carrier frequency for scanning and this is modulated by the change in the light value of the picture.

Suppose that we have a test picture of the form shown in Fig. 17, the black and white bars representing the finest detail

scanning the whole picture area twice per frame period. This means an effective scan from the flicker point of view of 50 times a second, although the effective scan for detail is only equivalent to 25 times per sec. The equivalent condition for a cinema film is that 24 pictures per sec. are shown, but flicker is avoided by an additional shutter which makes the effective pictures shown 48 per sec., and the 441 scan lines gives a picture detail which corresponds with that of the ordinary 16 mm. film.

The maximum band-width required will be up to 2·8 Mc/s, but the size of spot will be dependent upon the screen size. Thus with the normal screen of 10 $\times$ 8 cm. on the receiver, the spot will be of the order of 0·2 mm.

CHAPTER III

AMPLITUDE, PHASE, AND FREQUENCY MODULATION OF AN H.F. CARRIER

COMMUNICATION systems in general are associated with the transmission and reception of waves whose complexity of form and frequency depend upon the type of intelligence being communicated, and we have already seen that the effective frequency band of this intelligence varies from a few cycles per second up to megacycles per second.

One may conceive any communication system as having two essential features, each complementary to the other.

(1) A means of transferring energy from transmitter to receiver; thus we have sound waves in the case of speech between individuals; line currents in telegraphy and telephony; and ether waves in the case of wireless.

(2) A means of modulating this energy to conform to the desired intelligence, called the "signal." *

Systems may for convenience be divided into three classes:

(1) Non-carrier systems.
(2) Carrier systems.
(3) Suppressed carrier systems.

A non-carrier system is one in which energy only appears with the signal. If, however, a current traverses the link between transmitter and receiver even during the period of quiescence, we have what is known as a carrier system, this carrier current being modified in character by the signal.

And we may have yet another case, where, although no carrier traverses the link between transmitter and receiver, it originally existed in some form, but has been suppressed in the main transmission link and will again be incorporated at the receiver, such a system being known as a suppressed-carrier system.

* Throughout this chapter we shall refer to the modulation component as the "signal."

It may help to discuss, first, simple line telegraphy, considering it from perhaps a rather unusual point of view. A Morse telegraph signal consists of three elements, dot, dash and space, but in the only systems we shall consider, the dot and dash are both formed by pulses of current in the same direction but of different duration, and will both be classed under the term "mark." One element of a signal will, therefore, consist of a mark and a space, any one signal element of mark and space virtually comprising a square waveform.

There are two possible ways of transmitting such an element illustrated in Fig. 18 (a and b). The signal can be made by a transmitter sending a current of amplitude "a," say, along

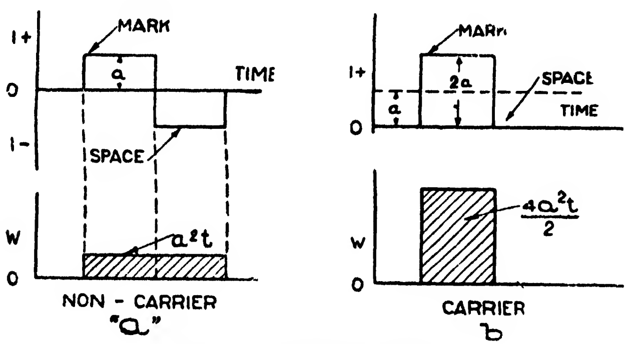


FIG. 18. Telegraph Signal.

a line, first in one direction, the mark, and then in the other, the space, giving an effective amplitude difference between mark and space of 2a.

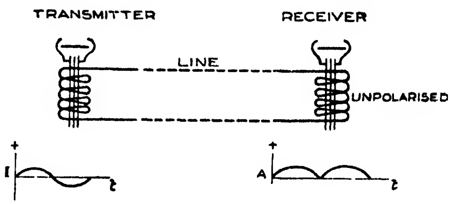
Considering the complete signal as a square A.C. waveform, we can say that the datum line of signal coincides with zero line of current in the system, and it is to be observed that a reversal of phase takes place at the half cycle of signal. This is an example of a non-carrier system, and this changing line current, which forms the signal, can be detected by a suitably designed receiver. This system is the usual "double current" telegraphy.

Carrier Systems

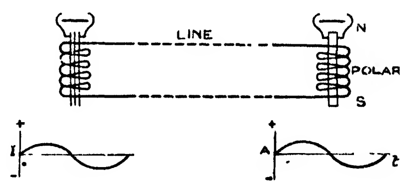
The second method is shown in Fig. 18b, and we may consider that a carrier current at least equal to "a" flows along the

line and is augmented and reduced (but not reversed) by the signal. If the carrier is equal to the signal amplitude, the line current on mark is 2a and on space zero. In this case the current always flows along the line in the same direction, the carrier forms a datum line for the signal wave, and is modulated to the shape of the signal.

Considering the power expended, it will be seen that in the first case it is half that of the second for the same amplitude signal, in spite of the fact that in the former current flows for both mark and space, and only during mark in the latter. The use of the carrier clearly involves an additional expenditure of energy which may be considerable particularly if the carrier amplitude exceeds the peak signal amplitude.



NON CARRIER SYSTEM
UNPOLARISED RECEIVER
FIG. 19.



NON CARRIER SYSTEM
POLARISED RECEIVER
FIG. 20.

A carrier current, or its equivalent is, however, often necessary because it aids detection at the receiver. For instance, compare the two line cases, just mentioned ; in the non-carrier example, for the receiver to detect the signal it has to be capable of following not only change of amplitude of current, but change of direction as well. On the other hand, if a carrier system is being used, the receiver has only to be able to detect change of amplitude and not change of direction. As an illustration, consider a line telephone; if an unpolarised receiver is used, it can only interpret change of amplitude of current and not change of direction, for any rise of current, no matter what its direction, attracts the diaphragm. Thus a signal of sine waveform would be received distorted to two half-sine waves, as shown in Fig. 19. If, however, the receiver

is polarised with permanent flux, giving an initial bias to the diaphragm, current in one direction augments the flux, and in the other direction decreases it, and in consequence a correct reproduction results as shown in Fig. 20. Had one of necessity to use the original type of detector it could be made to interpret both change of amplitude and direction of current by adding a D.C. carrier as shown in Fig. 21, although this carrier need not pass along the communication channel, but could be localised at the receiver circuit as shown in Fig. 22.

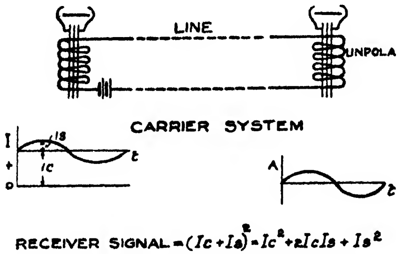


FIG. 21. D.C. Carrier System.

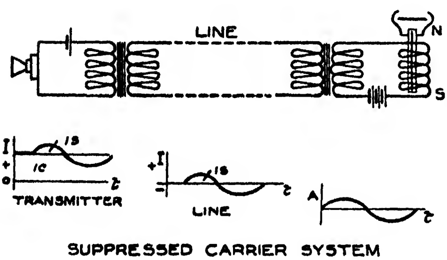


FIG. 22.

We cannot transmit L.F. signal waves directly by radio, and in consequence must use them to modulate an H.F. carrier, which is essential for the transmission of electromagnetic waves. Thus, instead of a D.C. carrier we have an A.C. carrier whose shape can be modulated by the signal, but in all cases we shall have an image wave (see examples, Figs. 25 and 26), and because of this image wave rectification is necessary at the receiver, for the purpose of extracting the signal.

Modulation of a High-frequency Carrier

A sinusoidal carrier current which is of constant amplitude, frequency, and phase, may be expressed thus:

$$i = A \sin (\omega_c t + \phi) \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

Where A = amplitude of the carrier current.
f_c = frequency of the carrier current.
$\omega_c = 2\pi f_c$.
ϕ = phase angle of wave, relative to some arbitary datum.

If we bring about a periodic change of either amplitude, frequency, or phase at a rate which is slow compared with the carrier frequency, the process is called modulation. The change must be made about a datum having the same value as the quantity (amplitude, frequency, or phase) had in the unmodulated carrier.

It must be appreciated that the effect of carrying out any of the above processes is to destroy the periodic character of the carrier. Thus in the case of an amplitude modulated wave shown graphically in Fig. 25, although the frequency remains constant throughout, each successive wave is no longer a sine waveshape, but distorted. Similarly with a frequency (or phase) modulated wave, shown in Fig. 38, although the amplitude is constant, the successive waves are no longer sinusoidal in character. This means a modulated carrier can no longer be expressed as a simple function but must be analysed into a representative series. This analysis of the carrier into a group of waves is a kind of extension to Fourier analysis, but simplified because it has been postulated that the modulation frequencies shall be small compared with the frequency of the carrier.

Amplitude Modulation

For instance, suppose we wish to modulate the amplitude of the carrier by a low-frequency sine wave $A_s \sin \omega_s t$. If $A_s = k_a A$ (where $k_a \times 100$ is the percentage modulation), then A in expression (1) becomes:

$$A\,(1 + k_a \sin \omega_s t) \qquad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

and it will be seen that this indicates a periodic amplitude change about the value (A) of the unmodulated amplitude. If we had merely changed A into $k_a \sin \omega_s t$, this would have indicated a change about a zero datum line.

By combining (1) and (2) and taking $\phi = 0°$, we obtain an expression for an amplitude-modulated carrier-current.

$$i = A\,(1 + k_a \sin \omega_s t) \sin \omega_c t \qquad . \quad . \quad . \quad . \quad (3)$$

This formula indicates that the synthesis wave consists of a high frequency carrier, constant in frequency but which is being varied in amplitude in accordance with the signal wave

about a datum displaced A from zero, Fig. 24 depicting the case of a sine-modulated carrier, 100% modulated.

Expression (3) can be expanded as follows :

$$i=A \sin \omega_c t+\frac{k_a A}{2} \cos (\omega_c-\omega_s)t-\frac{k_a A}{2} \cos (\omega_c+\omega_s)t \qquad (4)$$

This indicates that this same modulated carrier can be considered to be built up of a spectrum of constant-amplitude, constant-frequency waves consisting of the original carrier-frequency and two sets of high-frequency waves known as side bands, the amplitude of which will depend upon the value of k_a.

Thus if the signal is a sine wave as shown in Fig. 24, with k_a unity, there will be but two side-band waves, one each side of the carrier and differing in frequency from it by the signal frequency, the amplitude of each of these side-band waves being half that of the carrier.

If the signal is complex in character the amplitude changes produced will be such as to form an envelope (and reversed image) of the signal waveshape. But since any complex wave can be analysed into component frequencies, as was seen in the last chapter, each component frequency will produce with the carrier a pair of side-band waves of proportionate frequency and amplitude, as indicated by equation (4). This means the total high-frequency spectrum for a carrier modulated by a complex signal will comprise :—

(1) The original carrier.

(2) A band of H.F. waves obtained by taking the sum of carrier and signal frequencies, called the upper side band.

(3) A band of H.F. waves obtained by taking the difference of the carrier and signal frequencies, called the lower side band.

For instance a carrier modulated by the vowel sound "ah" shown in Chapter II, Fig. 6b, would appear as given in Fig. 23b, the spectrum of the vowel being as shown in Fig. 23a. The spectrum for the modulated carrier is seen to be obtained by moving the signal frequency spectrum along the frequency base by an amount equal to the value of the H.F. carrier ; and offsetting a reversed image spectrum on the lower side of f_c as if f_c was an origin for the signal frequencies, now

negative in effect. It will be noted that with such an amplitude-modulated spectrum the band-width is double that of the signal band, and is independent of the carrier frequency, or of the

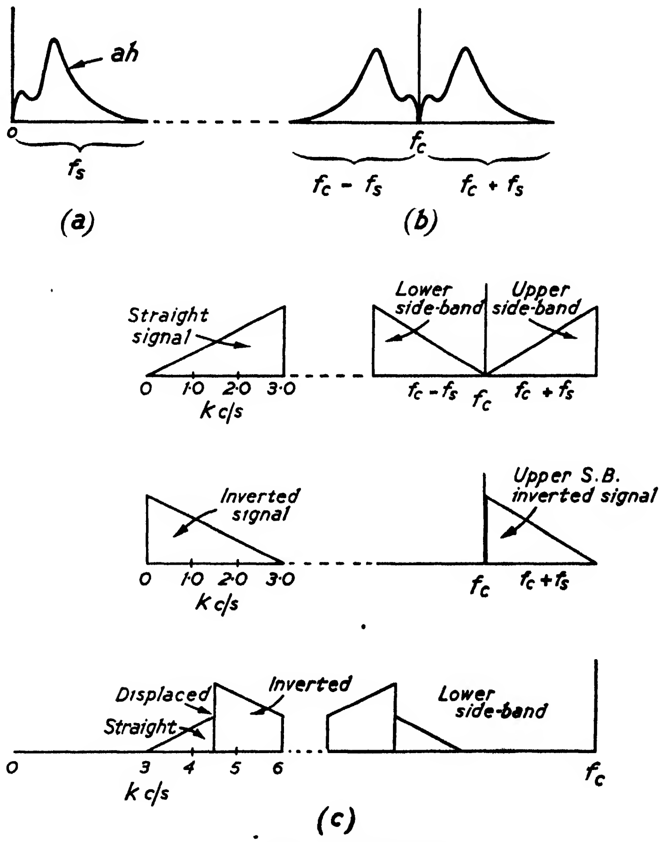


FIG. 23. Signal Representation.

modulation factor. The amplitude of the sidebands is, however, dependent upon the modulation factor.

A conventional method of indicating sidebands is now becoming usual and will be indicated here. The radio frequencies produced as a result of modulation are shown on a horizontal frequency-scale, but no attempt is made to show their amplitudes. Instead, the vertical ordinate is also a frequency scale showing the original signal frequencies. Thus

the ordinary double-side-band transmission will appear as in the top line of Fig. 23c, since in the upper side band the higher signal frequencies produce higher radio frequencies, whilst in the lower side band the reverse is the case.

Sometimes the signal band may be bodily displaced in frequency and split into bands which may, or may not, be inverted. Frequently only one sideband is transmitted. In Fig. 23c various examples are given to illustrate the convention.

Power Distribution in Amplitude-modulated Carrier

It is of considerable interest to discuss the distribution of power in such a built-up wave and to observe the effects of phase displacements. The question is easily studied with the aid of a mechanical synthesis machine* and in the discussion

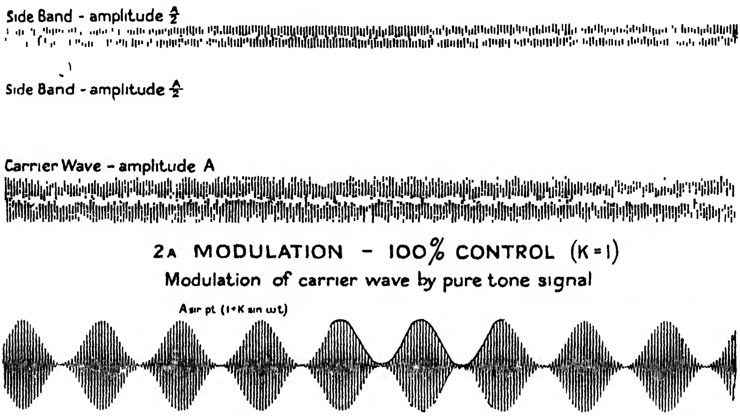


Fig. 24.

which follows examples of built-up modulated waves taken by such a machine developed by one of the authors will be used to illustrate the various points.

Graphical examples of a sine modulated carrier with full and half modulation are shown in Figs. 24 and 25. Considering Fig. 24, for 100% modulation, the distribution of power will be proportional to the square of the individual waves, and it is seen that there is twice as much power in the carrier

* Marconi Reviews, Nos. 9 and 10.

as there is in the two sidebands together. If the carrier is 50% modulated, as shown in Fig. 25, then the amplitude of each sideband will be one-quarter the carrier amplitude, and hence the total sideband power is now only one-eighth the carrier power.

If we over modulate, sideband power is produced in too great a proportion and a distorted wave results.

The power proportion of carrier to sidebands for a given percentage peak modulation varies with the shape of the signal

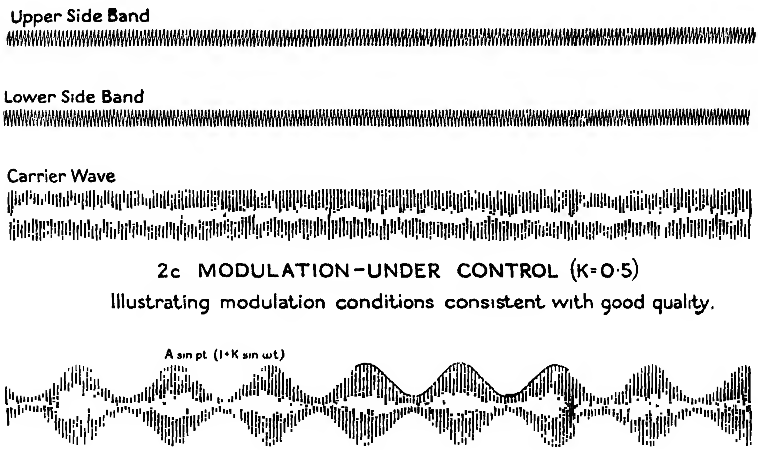


Fig. 25.

wave, the more peaky the wave form the less the power. It really comes to a question of wave area, and examples in Figs. 26 and 27 show this. Fig. 26 shows a carrier modulated by a wave approaching a square shape, as given by :—

$$F(t) = A \sin \omega t + \tfrac{1}{3} A \sin 3\,\omega t.$$

Such a wave will have two side-band waves each side of the carrier; one pair for the fundamental and one pair for the harmonic.

The power distribution in such a wave will be, approximately, 1 in carrier to ·556 in sidebands, showing that, because the wave is broader than a sine wave, for a similar peak its power content is greater than in the sine case.

Had the signal wave been more "peaky," say, as given by the expression :—

$$F_t = A \sin \omega t - \frac{1}{3}A \sin (\omega t + \phi)$$

where $\phi = 60°$,
although the same sideband frequencies are present, their phase is such as to increase the peak value of the wave and

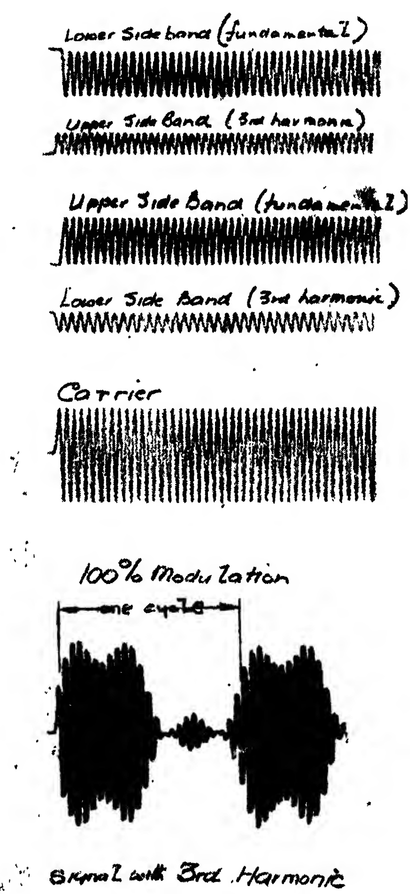


FIG. 26. Broad Signal Shape.

reduce its width. Since, however, it is assumed we are modulating the same amplitude of carrier, the sideband amplitudes must be less to avoid over-modulation. In this case the

power content of the sidebands to carrier is approximately ·312/1 for 100% modulation, as shown in Fig. 27.

It is obvious that when the signal wave is square, it will

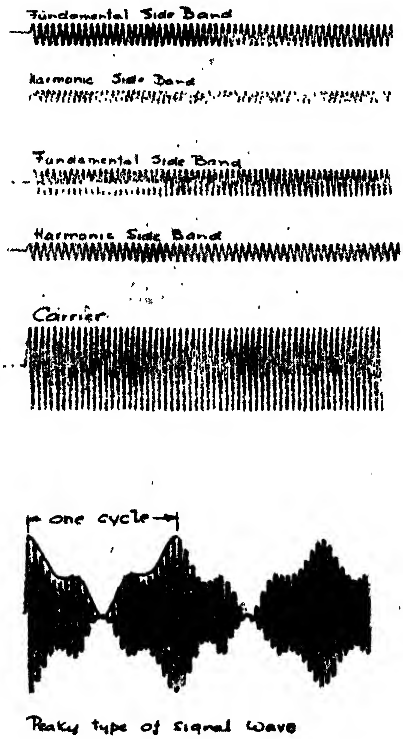


FIG. 27. Peaky Signal Shape.

then have the greatest power possible, and in this case the sideband power rises to equal that of the carrier.

Thus, under the most favourable conditions, there will always be as much power expended in the carrier as there is in the sidebands.

Actually, the transmission of square waves is rather a special case and control systems can usually be devised for it which are more efficient than the usual modulation circuits handling a changing waveform such as speech. We may regard the

$$F_t = A \sin \omega t - \frac{1}{3}A \sin (\omega t + \phi)$$

where $\phi = 60°$,

although the same sideband frequencies are present, their phase is such as to increase the peak value of the wave and

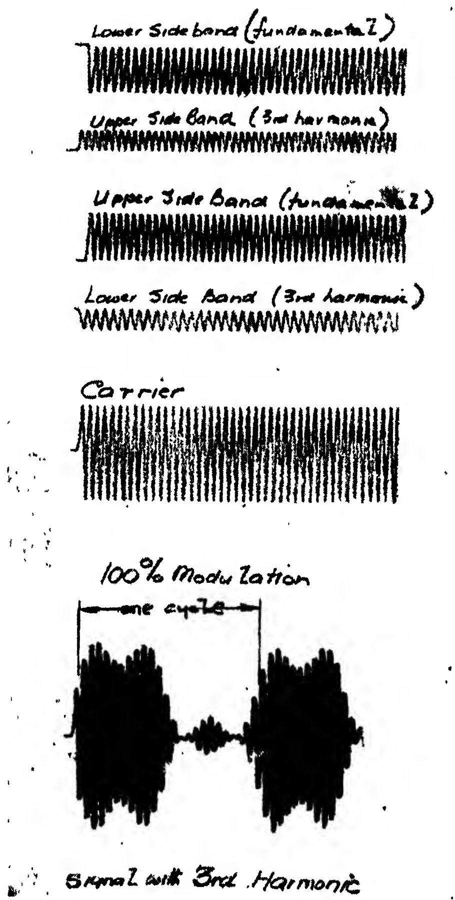


Fig. 26. Broad Signal Shape.

reduce its width. Since, however, it is assumed we are modulating the same amplitude of carrier, the sideband amplitudes must be less to avoid over-modulation. In this case the

power content of the sidebands to carrier is approximately ·312/1 for 100% modulation, as shown in Fig. 27.

It is obvious that when the signal wave is square, it will

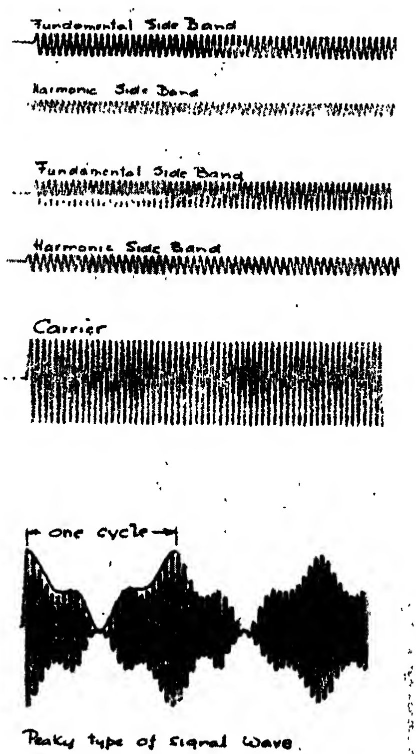


FIG. 27. Peaky Signal Shape.

then have the greatest power possible, and in this case the sideband power rises to equal that of the carrier.

Thus, under the most favourable conditions, there will always be as much power expended in the carrier as there is in the sidebands.

Actually, the transmission of square waves is rather a special case and control systems can usually be devised for it which are more efficient than the usual modulation circuits handling a changing waveform such as speech. We may regard the

sine wave as an average-shaped wave and its adoption for test purposes is therefore a good one.

Vector Analysis of Amplitude Modulated Wave

Although we have a spectrum of frequencies to deal with, ordinary vector analysis applied to an amplitude-modulated wave is a simple conception because the sideband waves are so disposed about the central carrier-frequency that each pair combines to form a frequency equal to that of the carrier. Consider the case of a sine-modulated carrier. We have to think of the addition of three rotating vectors, a carrier-vector C and two sideband vectors S_1 and S_2 of appropriate length,

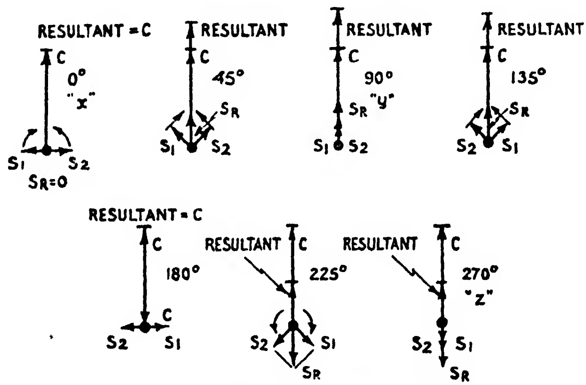


Fig. 28. Vectors for A.M. Carrier.

as shown in Fig. 28. Since at all moments one sideband is gaining on the carrier as much as the other is losing, we can, for convenience, consider the carrier vector stationary and the sideband vector S_1 rotating at a uniform speed of f_s cycles per second in a clockwise direction and S_2 rotating at a uniform speed of f_s cycles per second in an anti-clockwise direction. The addition of these three vectors will obviously result in a vector stationary in position (of the carrier frequency), but varying in amplitude at the modulation frequency, the cyclic variation of amplitude being obtained from the instantaneous resultant sideband vector S_R and the carrier vector C. Since this vector group will be rotating at constant speed it is clear that the resultant wave is one of constant frequency, varying in amplitude.

Such a method of visualising a modulated carrier provides a simple analysis, for the addition of any group of upper sideband vectors to its corresponding group of lower sideband vectors must always produce resultants which are exactly additive to or must be subtracted from the carrier directly.

Observe that the sideband vectors are rotating at constant angular velocity, once for every cycle of modulation and the angle traced out is equal to the proportion of modulation cycle completed, that is to say, since one cycle of modulation is represented by a complete revolution of each sideband vector, a quarter of a cycle of modulation is represented by angular positions of $\pm 90°$ of the vectors, half a cycle by $\pm 180°$ and so on. If we change the modulation frequency we change the rate of rotation of the sideband vectors, but one cycle of modulation will still be represented by 360° angular change, and in consequence there is always the same simple direct relationship between the modulation cycle and the angle traced out.

By plotting the vector amplitude on a time base we can show the variation of modulation envelope and if the ratio of carrier to modulation frequency is known we can also show the varying carrier that builds up this modulation envelope. It is to be noted that as long as the vectors are rotating at constant angular velocity no change of frequency is involved.

Carrier-suppression

In order to effect economy of working, the suppression of the carrier from the transmitter has been suggested, and we will discuss such a case.

If we suppress the carrier from the synthesis wave, the wave radiated and received is the beat produced by the side bands alone.

Fig. 29a shows the type of wave radiated and received from an original sine signal, and Fig. 29b the wave from a signal with wave-shape as shown in Fig. 26. We can trace out the original signal wave shape (shown by the thick line) and it will be seen that in each case the datum line of the signal wave now coincides with the zero line of high frequency. An interesting point to observe in the figure is the phase reversal occurring at every half cycle of modulating component, which

is analogous to the phase reversal of current in the line case. If reception of this carrierless spectrum is considered it will be clear that detection will not give the original signal because the detector is unable to interpret the meaning of the phase reversal at the half cycle, and in consequence the signal received will be of double frequency and distorted, an exactly analogous case to the unpolarised telephone. To obtain the original signal it is necessary to polarise the receiver with a high frequency carrier of the correct frequency and having a correct phase. The automatic phase reversal shown in Fig. 29a makes it clear how the addition of a polarising carrier at the receiver can build up the correct envelope shape. For the sidebands, each being displaced on the frequency band an equal amount above and below the carrier frequency, when added to each other form a beat wave of the carrier frequency; thus if a carrier is added so that it is in phase with the one-half of the modulated wave it raises the envelope up ; and because of the phase reversal, the carrier automatically assumes an antiphase condition with the second half cycle, the result being a corresponding reversal of envelope shape, and a correct reproduction of the original signal wave.

FIG. 29. Suppressed Carrier Wave.

It will be realised that it is extremely difficult to obtain two or more high frequencies sufficiently constant that they maintain the same phase relationship over a period of time (particularly at very high frequencies), and we must therefore discuss the effect of phase shifts on the resulting envelope. It should be pointed out in passing that one can only talk about phase relationship between waves of different frequencies at some point of reference, and in the particular case under discussion this is the phase of carrier at the instant of time when the side-band waves are in phase opposition; such a point of reference in the case of a sine modulated wave indicates the commencement of the signal envelope.

Effects of Phase Change in Re-introduced Carrier

Since only the side bands are received, the percentage modulation now refers to the ratio between re-introduced carrier and received side bands. The re-introduced carrier can easily be made large compared with the incoming signal so

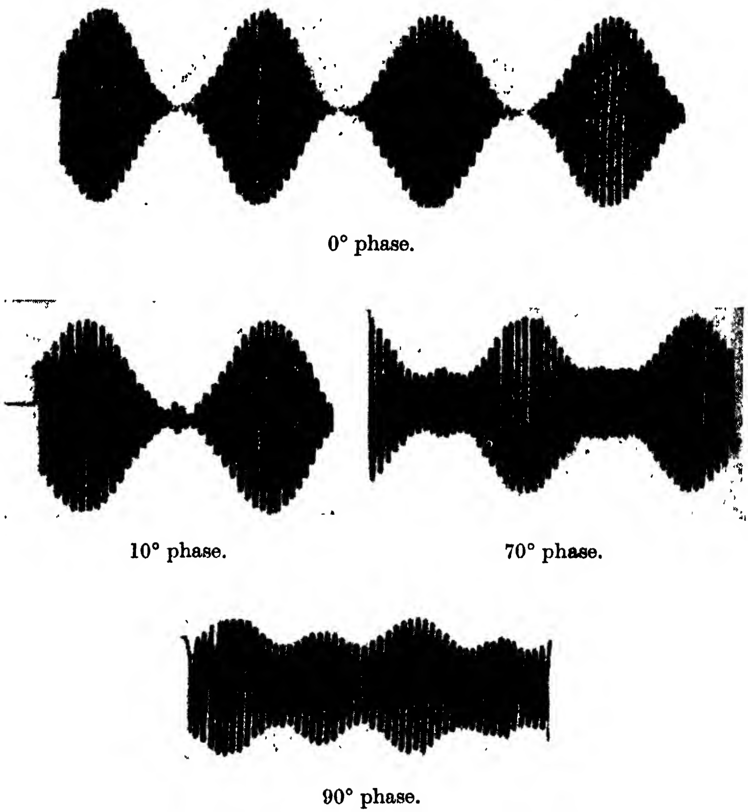


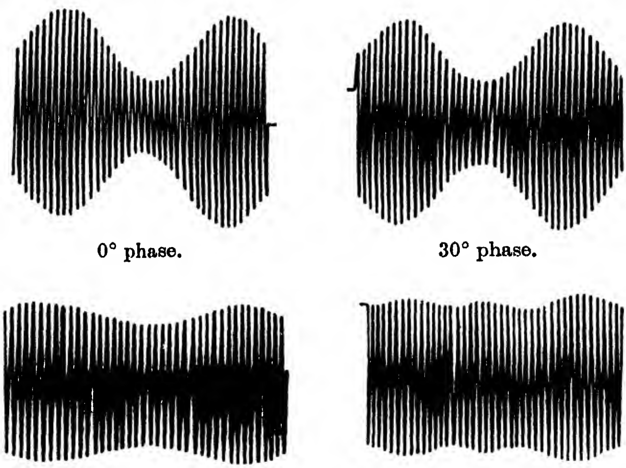
FIG. 30. Signal Distortion with Small Carrier.

that we are usually concerned with a very low percentage modulation. We will, however, discuss the general case.

Taking first the sine signal (sidebands as Fig. 29a), if we add a carrier of an equivalent size, *i.e.*, a modulation factor of unity, the resultant will be as shown in Fig. 30, where phase changes of carrier up to 90° are shown. These figures show

that considerable distortion occurs even with a small phase shift, and had the phase been continued beyond 90° similar envelopes would result, but tending towards a reversal when the phase reached 180°.

If the carrier amplitude is increased, or the percentage modulation reduced, there is less distortion apparent, as can be seen from Fig. 31. These curves show two very interesting features; first that as the phase is progressively increased to 90° there is a definite decrease in the effective amplitude



0° phase. 30° phase.

70° phase. 90° phase.

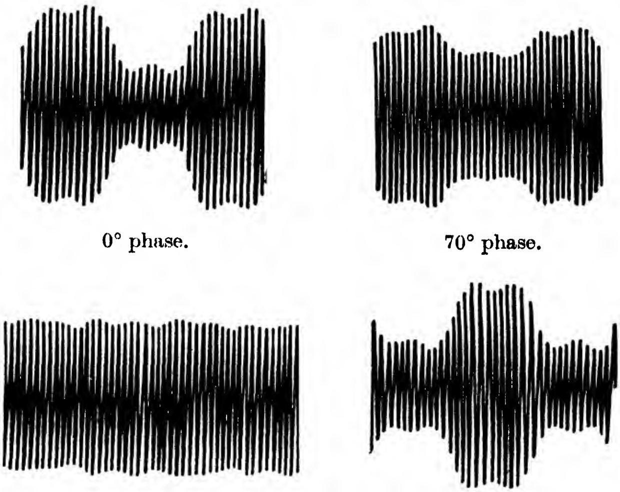
FIG. 31. 50% Mod. Small Signal Distortion with Large Carrier.

modulation; secondly, there is no change in the position of the signal envelope on the time base, only some distortion as the carrier approaches the quadrature condition.

A similar result is obtained with a more complex signal wave (Fig. 32), where phase shifts from 0° to 180° are given. This figure shows even more clearly the effects we have mentioned, namely, the reduction of modulation which accompanies the phase shifts up to the quadrature condition, the small amount of distortion, and that there is no phase change of signal in time, nor any relative phase between signal fundamental and harmonics. Also that with 180° phase shift the signal resumes its original depth of modulation, but reversed in sense. It can be imagined that if a signal envelope is to go

through a process of being " turned inside out ", there must be a transition stage where the envelope flattens out.

The fact that the amplitude changes decrease as the carrier phase approaches 90° does not mean that modulation by the signal has disappeared, but simply that it has changed its effect



0° phase. 70° phase.

90° phase. 180° phase.

FIG. 32. Complex Signal with Large Carrier.

on the carrier. Although it is not evident from the diagrams, a careful examination of such a wave when suitably enlarged would reveal that the lessening of amplitude modulation is accompanied by a growth of phase modulation (to be discussed later), this phase modulation reaching its maximum when the carrier and the sideband resultant are in quadrature.

Vector Analysis of Suppressed-carrier System

Consideration of the vector analysis shows how these results come about.

If now we consider a carrier having a phase shift of say 90° it means that the sum of the sideband vectors must not be added directly to the carrier, but in quadrature, as shown in Fig. 33 ; because of this, if the carrier vector is large compared with the sideband vector (as it is for shallow modulation) the resultant amplitude of these two vectors added in quadrature

will always be approximately the same, for a small vector added in quadrature to a large vector gives a resultant of almost the same amplitude as the larger vector, as shown in Fig. 33.

Considering the rotation of this group of vectors, our carrier vector is the reference point, and it is clear that the vector resulting from the carrier and the sidebands no longer maintains a constant position relative to the carrier, but it is at one time leading it, and at one time lagging behind it ; in fact there is a cyclic variation of phase at the modulation rate, and

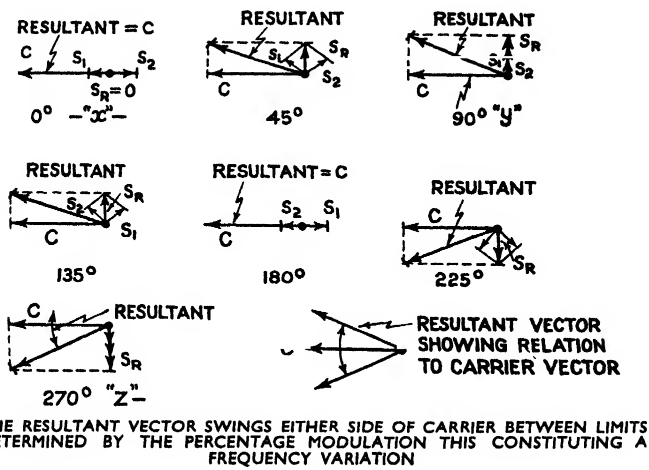


FIG. 33. Vector Diagram for Re-introduced Carrier Misphased.

we therefore come to the conclusion that the resulting wave is no longer constant in frequency but varies cyclicly at the modulation frequency, this frequency variation being accompanied by a small cyclic change of amplitude.

Phase-shifting the carrier of an ordinary amplitude-modulated wave has the effect, therefore, of reducing the amplitude modulation and introducing a phase or frequency modulation ; and with a carrier phase-shift approaching quadrature, the amplitude modulation may almost disappear, although a change of frequency will have been introduced which may become fairly considerable.

If a carrier is introduced of a different frequency it will beat

with the sideband spectrum and there will only be transient conditions where the envelope is reproduced correctly.

It would appear, therefore, that to prevent loss of amplitude modulation we must hold the carrier phase to within fractions of a cycle. If no pilot carrier is sent, this means the frequency stability of both transmitting and receiving oscillators has to be controllable within such limits, an impossibility even with the very high precision that present-day technique can offer. We are therefore led to the conclusion that a system suppressing the carrier alone is not a practical proposition.

Single Sideband Working

To overcome the difficulty of working a suppressed-carrier system a compromise has been effected by adopting what is known as "single sideband working," and this system is of considerable interest, for it not only has the power economy of a suppressed-carrier system, but the overall frequency spectrum radiated is reduced to half, a considerable advantage in itself in these days of a "crowded" ether. The system has certain limitations, as will be seen later, and suppressing one sideband effects no power economy in itself, since by doing so the effective signal strength is correspondingly reduced.

Consideration of the original high-frequency spectrum shows that either sideband contains the signal components. Single sideband working, therefore, consists of suppressing at the transmitter not only the carrier but one sideband as well, and thus one group of waves is transmitted having the frequency of carrier plus (or minus) the signal component frequencies; but it might be pointed out that by the suppression of the carrier and a sideband as well we have lost the reference frequency to which the remaining sideband must be added at the receiver. Suppression of the carrier can be carried out by a balanced circuit, as described in Chapter XVI, the sidebands being separated by any convenient form of filter circuit. As in the case of suppressed-carrier working, to reproduce the signal at the receiver it is necessary to add a polarising wave of the same frequency as that suppressed, since detection of the sidebands alone does not result in the extraction of the signal component.

The theory of action is that because the carrier differs from the waves of the side band by amount of the signal components, the beating of the group with a similar carrier produces a synthesis wave whose envelope has the component frequencies of the signal.

Thus, in the case of a sine signal of frequency f_s, modulating a carrier of frequency f_c, two sideband waves are produced (f_c+f_s) and (f_c-f_s). Thus, if the upper sideband is being used it will consist of a frequency (f_c+f_s). The addition of a carrier frequency f_c at the receiver before detection produces a wave

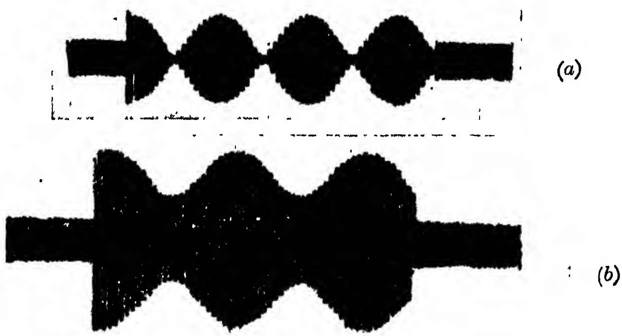


Fig. 34. S.S.B. Simple Signal.

whose beat envelope will have a frequency of $(f_c+f_s)-f_c$m namely f_s, the original signal frequency.

This simple arithmetic, however, is somewhat misleading, and it will be found that although f_s does appear as the chief frequency in the detector output, the original envelope shape is not reproduced, although under certain conditions the distortion can be reduced to a small amount.

Considering the simple case of a sine signal after suppression of the carrier and one sideband, the remaining sideband radiated is a constant amplitude H.F. wave (see Fig. 24) differing in frequency from the carrier by the signal frequency. If we add a local carrier of the correct frequency we produce an envelope of the original signal frequency. But if the carrier introduced at the receiver is too small, the waveshape will not be of the original sine form, but of half-sine form at the signal frequency as shown in Fig. 34a. Increase of carrier

amplitude leaves the signal envelope-frequency unchanged, of course, but the waveshape approaches more and more to the correct sine waveform as the carrier becomes very large compared with the sideband, as shown in Fig. 34b. This apparent change of envelope shape merely by an alteration of the relative amplitudes of the waves beating together is not easy to understand physically. It can be shown mathematically to be logical, and can be summarised briefly thus. The expression for the rectified envelope of such a wave can be shown to consist of the product of two simple functions, *i.e.*, a complex waveshape. One of these functions, however, reverts to unity when the amplitudes of the beating waves are very widely different in amplitude, and the rectified envelope can then be expressed in terms of a simple function.

The same effects are, of course, obtained with complex signals, and with single sideband work it is essential therefore to use a very large amplitude carrier at the receiver. Other effects can best be discussed by considering a more complex signal.

Complex Signals and the Effect of Phase Changes in Single Sideband System

Reverting for a moment to the sine signal, since the side band consists of but a single H.F. wave it is clear that the same envelope will be produced whatever may be the phase of carrier introduced. It is not easy to see with such a simple case the relative phase of signal envelope on the time base, but a little thought would show that the envelope time phase will change with the H.F. phase change. Thus, if carrier and sideband had zero phase, they would add at that instant to produce the peak of the signal, whereas if the carrier had at that time a lag of, say, 60°, the two waves would come into phase $\frac{1}{6}$th of a signal cycle later to produce the peak of the signal.

Let us consider a complex signal, whose envelope is that given previously in Chap. II Fig. 8 (top) and which comprises a fundamental wave and a third harmonic with zero phase. The modulated carrier wave-shape will be as Fig. 26.

If we suppress the carrier and one sideband, what is radiated and received will be one sideband only. If we assume the "lower" is radiated, the envelope of this radiated wave, shown

in Fig. 35, appears to have no obvious relationship to the original signal.

If carefully examined, however, it is found that the envelope actually gives the difference frequencies of the wave making up the signal. Thus in the present case of a signal with frequencies 1 and 3, the difference 2, shown by the envelope.

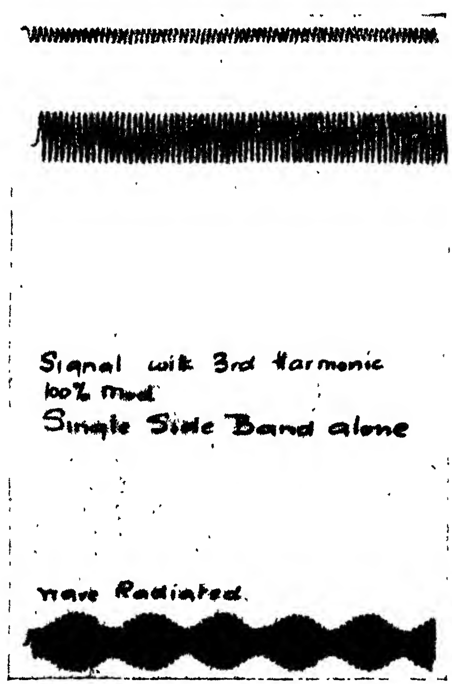
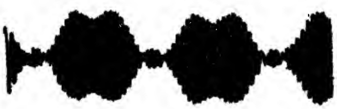


FIG. 35. S.S.B. Wave Radiated.

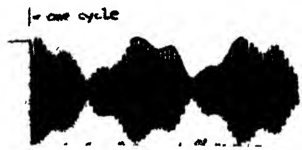
It is for this reason that with a pure tone signal, since there is only one signal frequency, the difference frequency is zero, and hence the single sideband signal becomes a continuous wave.

The introduction of the correct carrier frequency gives us the correct signal waveshape as seen from Fig. 36a, although some distortion is apparent by the "peaky" character at the minimum of the signal due to the introduced carrier not being of large enough amplitude. If the phase of carrier is changed,

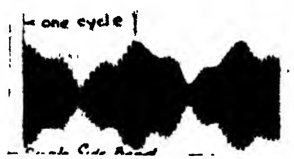
the signal wave-shape changes as shown by Figs. 36b, 36c, and had the phase been 180° the envelope would have been reversed on that of Fig. 36a. At first sight these various envelopes appear to represent different wave spectra, but if each envelope is examined it will be found to contain the same signal frequency components, namely one and three, and having the same amplitude relationship, as can be seen by comparing these envelopes with those in Fig. 8. It is only the relative phases of the signal components that have altered. Thus with a signal wave rich in harmonics, not only will the phase of the fundamental be shifted, but the phase of each individual



(a) 0° phase.



(b) 45° phase.



(c) 90° phase.

FIG. 36. S.S.B. Synthesis Wave with added Carrier.

harmonic shifted a different amount, and if one considers the general case it is observed that the phase shift of the different signal component frequencies is directly proportional to those harmonic ratios.

Because the shape of the resulting envelope depends upon the relative phases of the different components of the signal it follows that in single side-band working the envelope shape alters considerably with a phase shift of carrier. If, however, such a wave is being received aurally, the intelligence conveyed by the ear to the brain will be exactly the same whatever envelope shape is made by the combination, for the ear is only capable of interpreting frequencies and amplitudes, and

not phases. It interprets frequencies accurately, amplitudes indifferently, and phases not at all, this statement being true for all waves which are not transient in character.

If the added carrier is altered in frequency, the pitch of the fundamental will alter, and since the difference frequency to the harmonics changes, the signal harmonic frequencies will alter relative to the signal fundamental, but it should be observed that because the signal fundamental has the smallest difference of frequency to the carrier, a small change of carrier frequency makes a larger percentage change in the fundamental of the signal than it does in the harmonics. Now if one considers a speech signal, the fundamental conveys mostly pitch and the harmonics mostly character, and because the harmonics are changed comparatively slowly compared with the fundamental, one can allow quite a remarkable amount of carrier frequency change before speech becomes unintelligible, the most marked alteration being the rapid change of voice pitch. The above statements are true for speech but not for the transmission of music, for in the latter case much of the pleasure (or otherwise) lies in the combination tones of the various frequencies produced, and if these are upset the balance of consonance and dissonance is upset.

The general inference is, then, that single sideband working offers a wide field of development for certain classes of traffic, but that field will be limited to systems employing aural reception, unless distortion of envelope can be permitted, frequency characteristic of circuits improved or synchronising systems adopted.

Actually no intelligence can be conveyed by a single sideband system unless :

(*a*) The nature of the intelligence being conveyed is known to the receiving operator, or :

(*b*) The exact carrier frequency originally suppressed by the transmitter is known at the receiving station.

As regards (*a*), in the case of ordinary telephony, for instance, the receiving operator knows that he has to receive intelligible speech and in consequence automatically adjusts the system until this is obtained. But suppose the intelligence required is a pure tone, then since the single sideband is now but a

single frequency of constant amplitude, the receiver can in no way collect the original intelligence accurately.

In connection with (*b*), although we may and do know the alleged frequency of transmitted carrier, it must be remembered that we have been discussing fractions of a cycle per second. And since frequency constancy is a matter of parts of a million at the best, even with stations having so-called constant frequency sources, we never know the exact frequency from moment to moment, without the transmission of a pilot carrier signal.

Phase and Frequency Modulation

As indicated in a previous section of this chapter, alternative methods of modulation involve phase, or frequency, changes

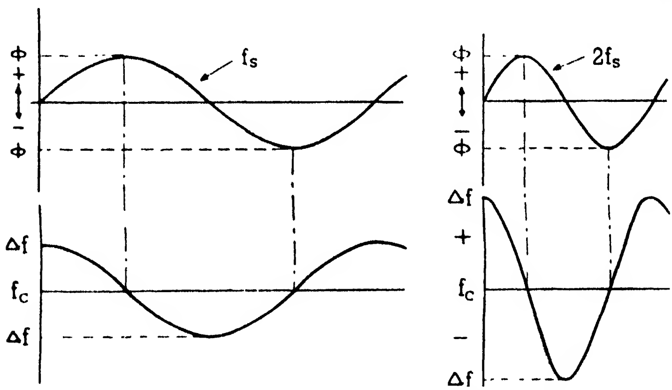


FIG. 37. Phase and Frequency Modulation.

of the carrier wave, instead of amplitude. As with the latter, either method involves a distortion of carrier waveshape and in consequence the production of side-band waves. In phase modulation the signal produces a change in the phase of the carrier, referred to an arbitary datum, which is proportional to the instantaneous amplitude of the signal. With frequency modulation, the signal produces a deviation in frequency of the carrier which is proportional to the instantaneous amplitude of the signal.

If we consider modulation by a sinusoidal signal, both systems are essentially the same in effect, because a sinusoidal

phase displacement represents a frequency deviation proportional to the rate of change of slope of the phase-time curve, and this is still sinusoidal as indicated in Fig. 37, which shows the phase displacement time curve and the resultant frequency deviation time curve beneath. It will be observed from Fig. 37, where curves for two different signal frequencies are given, that at points of maximum phase displacement the rate of change of phase is zero and in consequence the frequency is that of the unmodulated carrier; whereas at points where the phase displacement is changing most rapidly, the frequency deviates from the carrier by an amount which is proportional to this rate of change of signal, the higher frequency giving a greater deviation frequency. Thus, with a phase-modulated carrier, although the phase displacement is a constant for a given amplitude of signal, the rate of change of phase and in consequence the deviation frequency increases with the signal frequency, the ratio $\frac{\Delta f}{f_s}$ remaining a constant.

Although there is a similarity between a phase and a frequency-modulated wave for an applied sinusoidal wave, this is only because the differential of a sine wave is still sinusoidal in form; for other types of signals there is a distinct difference, as will be seen later, but it is a simple matter to convert from one form to the other.

Phase Modulation

Consider a carrier wave whose amplitude A and frequency f_c remain constant, but whose phase is varied sinusoidally about a mean value ϕ by an amount $\pm k_p\phi$ where k_p is a factor proportional to the signal amplitude, and $k_p\phi = m_p$ is known as the Modulation Index, k_p becoming unity when the applied signal amplitude is a maximum. If ϕ_0 is the maximum phase change, when k_p is unity a frequency deviation Δf is produced equal to $m_p f_s$ c/s, or:

$$m_p = k_p\phi_o = \frac{\Delta f}{f_s}$$

It should be remarked that with phase (and frequency) modulation there is no rigid upper limit of modulation due to the carrier, as there is with amplitude modulation, although

circuit considerations do impose an eventual limit. It is found that with phase (and frequency) modulated circuits we can provide for a signal amplitude ratio of some 100/1, *i.e.*, 10,000 power ratio, or 40 db, as compared with a signal amplitude ratio of about 20/1, *i.e.*, 400/1 power ratio, or 30 db, with an amplitude modulated carrier.

A phase modulated wave would be depicted as in Fig. 38 which shows the synthesis wave of constant amplitude, and

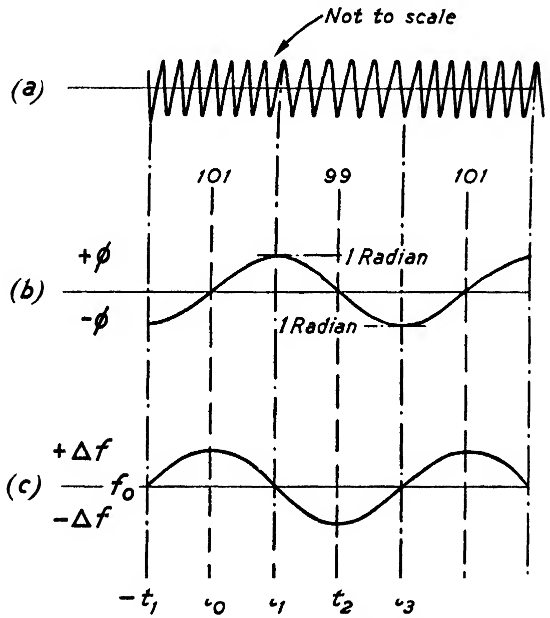


FIG. 38. Phase Modulation.

the relationship of phase change and frequency deviation resulting, for a given sine modulating signal. By following reasoning similar to the amplitude modulated case we have :

$$i = A \sin\left\{\omega_c t + \phi\,(1 + k_p \sin \omega_s t)\right\}$$

$$= A \sin\left(\omega_c t + \phi + k_p\phi \sin \omega_s t\right) \qquad (5)$$

Let $m_p = k_p\phi$ and assume the carrier phase-angle is 0°. Then it can be shown that :

$$i = A\Big[J_o(m_p)\sin\omega_c t + J_1(m_p)\Big\{\sin(\omega_c+\omega_s)t - \sin(\omega_c-\omega_s)t\Big\}$$
$$+J_2(m_p)\Big\{\sin(\omega_c+2\omega_s)t + \sin(\omega_c-2\omega_s)t\Big\}\ldots\ldots$$
$$+J_n(m_p)\Big\{\sin(\omega_c+n\omega_s)t - \sin(\omega_c-n\omega_s)t\Big\} \quad . \quad (6)$$

Where $J\ (m_p)\ J_1(m_p)$ are Bessel functions with argument m_p.

Examination of Equation 6 shows that the sine-signal modulated wave comprises a carrier and two groups of side-band waves forming an infinite series thus:

$$f_c \pm f_s \cdot f_c \pm 2f_s \cdot f_c \pm 3f_s \cdot f_c \pm nf_s \ldots$$

Unlike the amplitude modulated carrier, which has but one

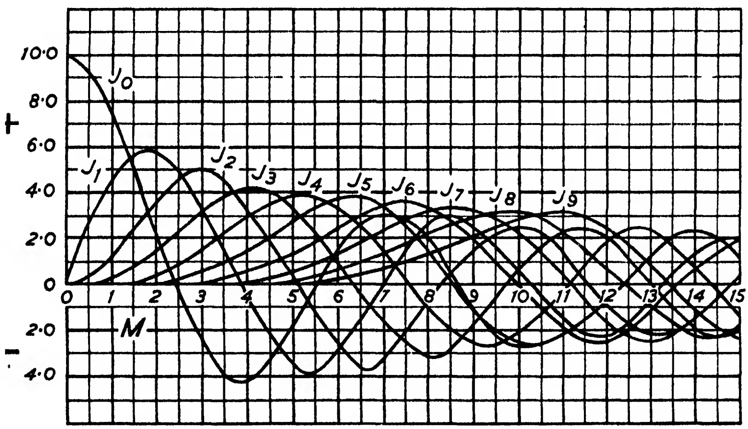


FIG. 39. Phase Modulation; Sideband Amplitudes.

pair for each sine modulation component, we have an infinite spectrum, although, of course, only those with amplitudes above a certain minimum (usually of the order of 1/100th of the original carrier) are of importance.

The amplitude of these sidebands, and that of the carrier during modulation, are not in an ordered sequence, but vary relatively to one another as signal amplitude and frequency changes.

Further, the amplitude of these sidebands and that of the carrier have no simple relationship as they have in amplitude modulation, but vary rapidly as either the signal amplitude or

frequency is changed. The relationship of carrier and the first side band amplitudes to modulation index is shown in Fig. 39, which sets out the Bessel function coefficients plotted against the argument m_p, up to a maximum value of m_p of 15.

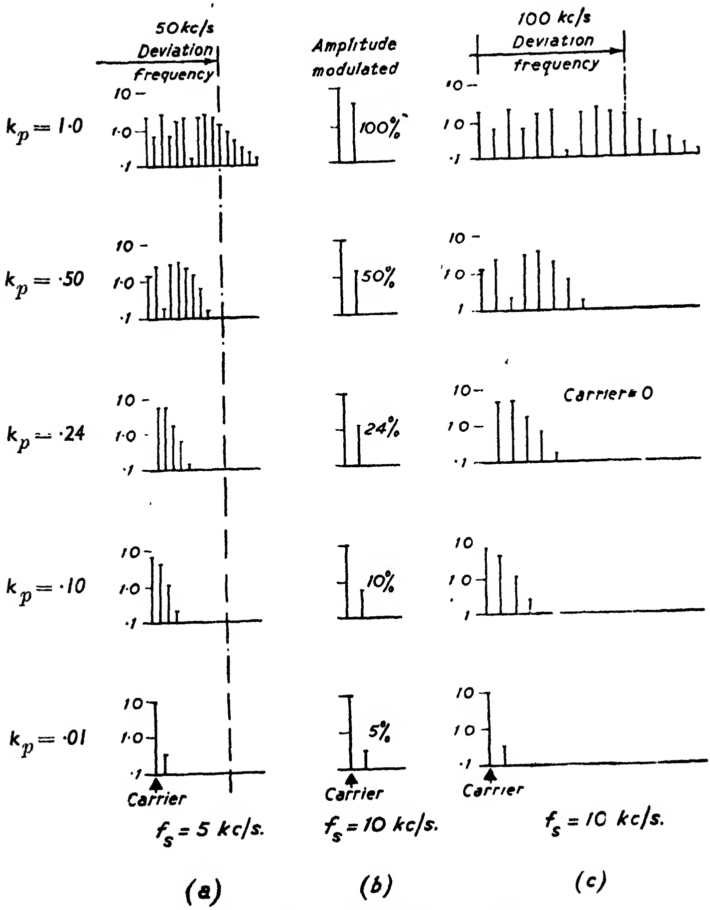


FIG. 40. Phase Modulation Spectra.

Thus, observe that when m_p is 2·4 the carrier amplitude is reduced to zero, and again with values of 5·6, 8·7, etc.

Another difference from the amplitude-modulated carrier is that the phase of sidebands (as previously defined) is not zero, but each pair alternates between zero and quadrature.

The odd frequency pair are in quadrature with the carrier, and the even pair in line, *i.e.*, one in phase and one in antiphase. If reference is made to a previous example of quadrature phase of a single pair of sidebands (see page 60), it was seen that the resultant had not only a frequency variation but a small amplitude variation as well; the introduction of this infinite series with phase alternating between quadrature and in line is to level out this amplitude difference.

Fig. 40 (a and c) shows the frequency spectra (with amplitudes drawn to a log scale) of a carrier, phase-modulated by sine frequencies of 5 and 10 kc/s., the maximum phase displacement being 10 radians. Different signal amplitudes over a 100/1 range are shown. Only the sidebands on one side of the carrier are shown and an amplitude-modulated spectrum is shown for comparison.

The number and amplitudes of the sideband waves are independent of the frequency of the carrier, and can be obtained relatively from Fig. 39, as can the amplitude of the carrier, for different values of k_p. Thus, if we have a maximum amplitude applied, as given by $\phi = 10$ radians $k_p = 1{\cdot}0$ and hence $m_p = 10$ and it can be seen from Fig. 39 that the carrier amplitude is equal to $-2{\cdot}6$ and there will be some 15 sideband waves whose amplitudes can be obtained from the coefficient values in Fig. 39. These values are all shown positive in Fig. 40a and c. Whereas if we consider an applied signal of amplitude ·24 of the maximum, shown at 2·4 (Fig. 39), the carrier has now disappeared, and there are now only five sideband waves, the spectrum being shown in Fig. 40a and c, for modulation index of ·24.

Thus, as we change the signal amplitude, from maximum to minimum, instead of a simple reduction in the amplitude of the sideband waves as with the amplitude modulated carrier, we obtain a change in the spectrum width. Thus at full modulation the band-width is greater than the deviation frequency, and progressively decreases until at very low level it is similar to the amplitude-modulated carrier with but a single pair of sideband waves.

Considered in terms of phase shift, when ϕ is very much less than 1, the deviation frequency is much less than the modulation frequency, and the band-width is then about twice the latter.

But when ϕ is much greater than unity, the deviation frequency is much greater than the modulation frequency and the band-width is then greater than twice the deviation frequency.

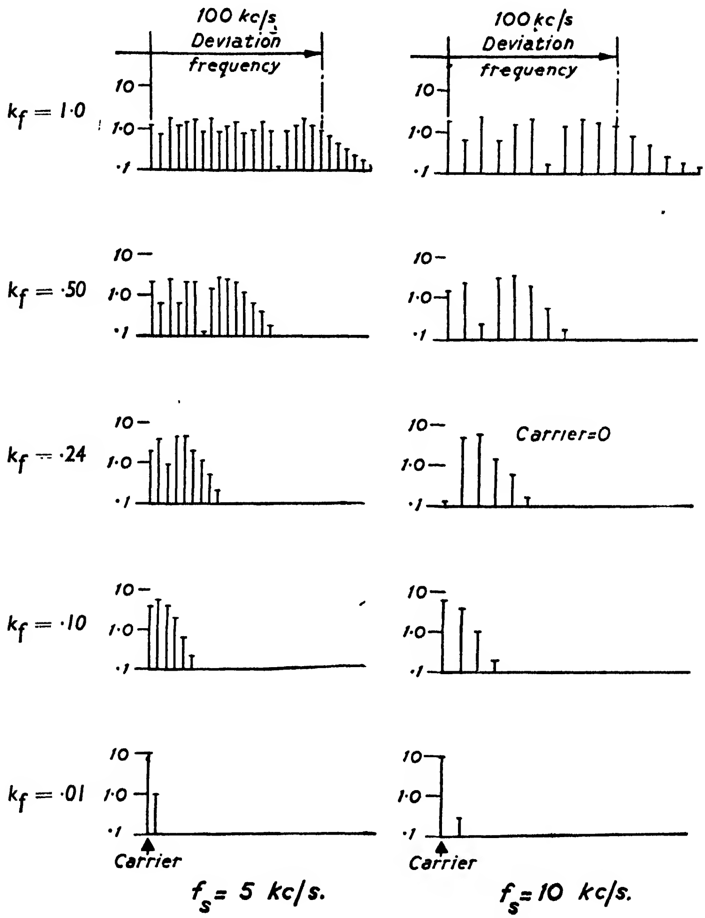


FIG. 41. Frequency Modulation Spectra.

With ϕ unity, $\triangle f$ equals f_s and the band-width is then some six times either $\triangle f$ or f_s.

Frequency Modulation

Following similar reasoning it can be shown that a frequency-modulated current wave may be expressed :

$$i=A\Big[J_o(m_f)\sin\omega_c t+J_1(m_f)\Big\{\sin(\omega_c+\omega_s)t-\sin(\omega_c-\omega_s)t\Big\}$$
$$+J_2(m_f)\Big\{\sin(\omega_c+2\omega_s)t+\sin(\omega_c-2\omega_s)t\Big\}$$
$$\ldots\ldots+J_n(m_f)\Big\{\sin(\omega+n\omega_s)t-\sin(\omega_c-n\omega_f)t\Big\}\Big].\quad(7)$$

where $m_f = k_f \dfrac{\Delta f}{f_s}$

k_f becoming unity for full signal amplitude. Comparing this with expression (6) it is clear the resultant spectrum for a frequency-modulated wave will be very similar in form to that for a phase-modulated wave, the only difference being in the value of the coefficients m_p and m_f. At the signal frequency for which these coefficients are equal the two spectra will be identical (compare the 10 kc/s signal), but any change of frequency would produce a difference. We can say that as with phase modulation, the band-width is in general determined by whether Δf or f_s is the greater, and where one frequency is much greater than the other, the band-width is rather more than twice the greater frequency, the frequency spectra for frequency-modulated carriers being shown in Fig. 41 for comparison with those of the phase-modulated case.

Vector Analysis of Phase and Frequency Modulation

With amplitude modulation we were interested in a synthesis wave of constant frequency and varying amplitude. The vector representation of such a wave is explicit and simple because the length of the vector shows directly the instantaneous amplitude, whilst the angular position from a given datum shows the phase relative to that datum. Because we are dealing with a constant frequency, the constant angular rotation of the vector scarcely enters into the argument. The use of this same vector convention to express a phase or frequency-modulated carrier becomes more difficult, because we have a carrier whose frequency is varied. Thus instead of a vector rotating at constant speed but varying in amplitude per modulation period, phase or frequency modulation will be represented by a constant amplitude vector having a cyclic

variation of angular position from a given datum, and a variation of angular velocity per modulation period, a rather more difficult conception.

Consider a carrier of, say, 100 c/s, phase modulated by a sine signal of frequency of 1 c/s to a maximum phase variation of ± 1 radian. If we consider as a datum vector, one rotating anti-clockwise at 100 c/s, then a vector representing the modulated carrier will, for half the modulation cycle, gain in speed on the datum, and for the other half-cycle, lose in speed on the datum, reaching a maximum phase angle, leading or lagging, of 1 radian. If we adopt the previous convention of holding the datum vector stationary, then the phase-modulated vector will swing sinusoidally to and fro either side of this datum to a maximum angular position of ± 1 radian, a complete cycle of swing occupying 1 second, the signal modulation time. Fig. 42c shows the vector diagram, and Fig. 38 correlates this to the carrier wave changes.

Thus from a time t_0, when the vector has zero phase, it swings anti-clockwise to t_1, stops, reverses in direction and swings clockwise through the datum vector position at a time t_2, when it reaches its maximum velocity, slows up to t_3, stops, and passes through the datum vector again at t_4 at a maximum velocity. Clearly when the vector is travelling anti-clockwise, the frequency is greater than the carrier, and less when the vector is travelling clockwise, the deviation frequency being proportional to the velocity of the vector. At the ends of travel, when the vector is stationary, the wave radiated is that of the unmodulated carrier, 100 c/s. In the example given the vector swings 4 radians in a complete cycle, but as this is not at uniform speed, but sinusoidal, the deviation frequency is $\frac{4\pi}{2}$, = one cycle per second above and below the carrier frequency, *i.e.*, 99 and 101 c/s, as shown in Fig. 38a and c.

If we halve the signal amplitude, the angle of travel and the frequency swing are both halved. If the signal frequency is raised or lowered, then, to cover the same angular movement $\pm\phi_0$, the deviation frequency, or velocity of vector travel, must be raised or lowered in proportion, so the same vector diagram will be representative of a constant amplitude signal of any signal frequency, the difference being in the velocity of vector

as it passes the datum, which cannot be indicated except by a numeral.

Turning now to frequency modulation, let us consider a carrier of 100 c/s modulated by a sine signal of 1 c/s to a maximum deviation frequency of 1 c/s. Since these figures are as those just given for the phase-modulated wave, the

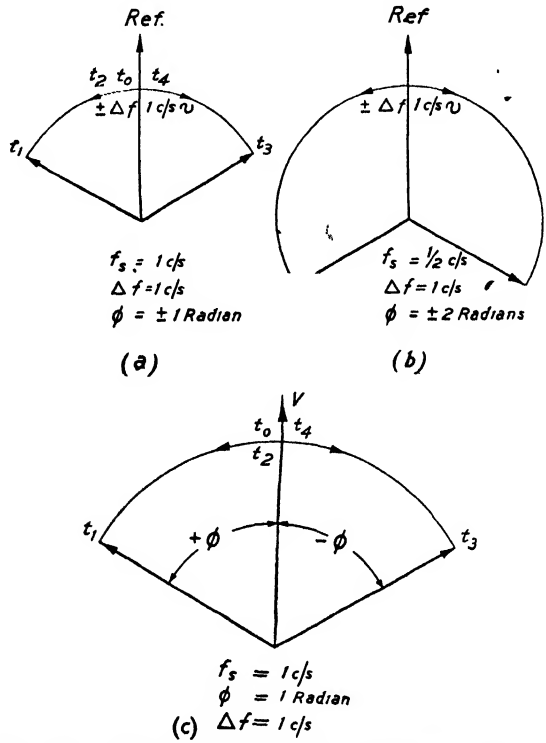


FIG. 42. Vectors for Frequency and Phase Modulation.

synthesis wave would appear as in Fig. 38a and c. As before, this wave will be represented vectorially by a swinging vector, whose maximum velocity as it passes the datum vector represents the deviation frequency, and whose total arc of travel is dependent not only upon the deviation frequency but upon the signal frequency as well, because the longer the time taken to carry out the swing, *i.e.*, the lower the modulation frequency, the greater will be the swing in either direction, In the present

case, the modulation frequency is 1 c/s, so the vector will have half a second to swing in either direction. If the velocity was uniform it would swing 180° in half a second, but since the vector is swinging sinusoidally and only reaches its maximum velocity at the centre of its swing, the vector will not trace out 180° but only $\frac{2}{\pi}$ of 180°, or 2 radians, as shown by Fig. 42a.

If we increase the signal frequency, keeping the same deviation frequency, then the arc of travel, or phase change, is decreased, since with the increased signal frequency the vector has less time to travel, its velocity being the same. Conversely, if the signal frequency is decreased the arc of travel is increased proportionally. (Fig. 42), showing vectors for the same deviation frequency and two signal frequencies, 42b being for a signal of half the frequency of 42a. It is seen therefore that the difference in vector diagram for the phase and frequency-modulated carrier is that, with the former, for a given signal amplitude the vector arc of travel is constant irrespective of the signal frequency, whereas in the frequency-modulated case the velocity is the same, but the arc of travel changes, increasing as the signal frequency decreases.

Signals other than Sinusoidal

It was mentioned earlier that although phase and frequency modulation are essentially the same for a sinusoidal signal, there is some difference when complex signals are applied and we can show the difference best by considering the modulation by a square signal shape.

Thus, consider the last example modulation of a 100 c/s carrier, by a signal of 1 c/s but of square shape. Taking the frequency-modulated carrier first, and assuming the deviation frequency to be as before, viz. 1 c/s, the modulated carrier and deviation frequency time curve would be as shown in Fig. 43a and b. That is from time t_1 the frequency is increased to 101 c/s, reduced suddenly at t_2 to 99 c/s, remaining at 99 c/s up to time t_3, and changing again suddenly back to 101 c/s, and so on. Thus at these times t_1, t_2, t_3, etc., there are transient conditions where the carrier changes and passes through the carrier value of 100 c/s. Vectorially this modulated wave will be represented by a vector which rotates at a *constant* speed

in either direction at 1 c/s, the transient carrier condition being shown by the stationary vector at the ends of the arc of travel. As before the total arc is dependent, not only upon the deviation

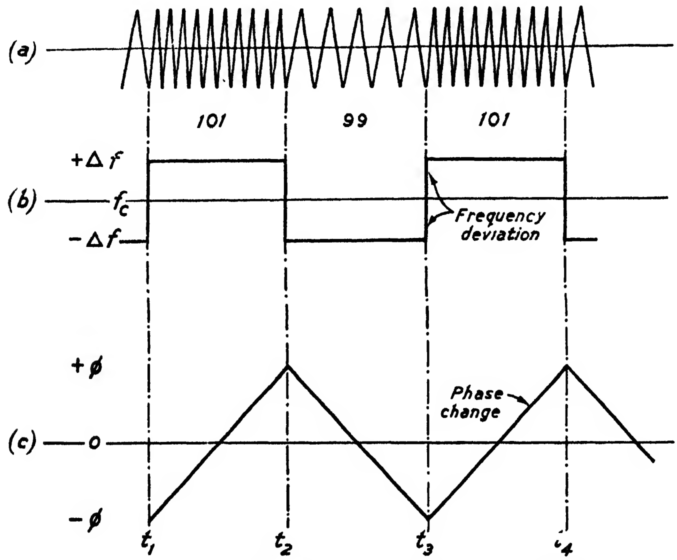


Fig. 43. Frequency Modulation—Square-wave Signal.

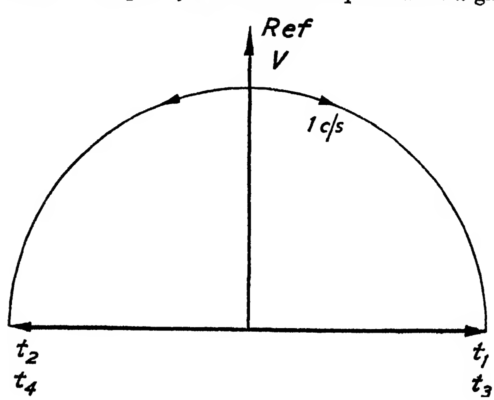


Fig. 44. Vector for Square-wave Signal.

frequency, but also upon the modulation cycle. Since it is travelling at a uniform speed of 1 c/s, and it has one second to complete the cycle, in this particular example it will trace out an arc of 180° as shown in Fig. 44.

Since the vector travels uniformly and not sinusoidally, the phase angle traced out is uniform as shown in Fig. 43c, from which it is seen that a square wave frequency modulation gives a triangular-shaped phase modulation. If now we consider a carrier of 100 c/s phase modulated by a square wave, say, to $\pm\frac{\pi}{2}$ radians, the result is as shown in Fig. 45, which indicates the corresponding frequency deviation at 45. Since the carrier is phase modulated by a square wave its frequency remains

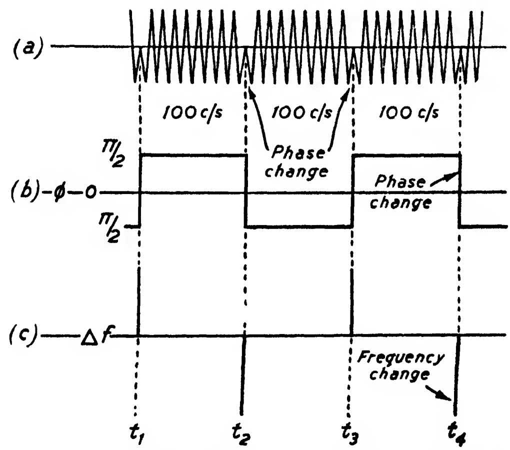


Fig. 45. Phase Modulation—Square-wave Signal.

constant, but the phase changes each half-cycle by an amount of $\frac{\pi}{2}$ radians from the reference position. Thus in Fig. 45, from the time t_1 to t_2 the frequency is 100 c/s but the phase is $\frac{\pi}{2}$ in advance of the reference, from t_2 to t_3 retarded by $\frac{\pi}{2}$, and from t_3 to t_4 in advance again, and so on, these discontinuities representing a momentary infinite change of frequency above and below the carrier as shown in Fig. 45c.

The vector diagram is similar to that for the frequency-modulated case and shown in Fig. 44. Except the vector, instead of travelling steadily " to and fro," now jumps suddenly from each extreme position where it has remained stationary

for half a cycle. Thus at the time t_1, Fig. 45, the vector swings suddenly from $-\frac{\pi}{2}$ to $+\frac{\pi}{2}$, *i.e.*, from t_1 to t_2, Fig. 44, remains stationary for the half-cycle t_1, t_2, Fig. 45, and then at the time t_2 jumps back again to the position $-\frac{\pi}{2}$, and so on. The vector is thus stationary throughout except for the transient changes each half-cycle which must be regarded as taking place at an infinite speed.

A Comparison of Modulation Systems

Amplitude modulation is the oldest method of transmitting telephony, etc., by radio and it is still used in 99% of such communications. Frequency and phase modulation have received a great deal of publicity in recent years, but it must be made clear that these systems have advantages over amplitude modulation only in a limited field. Of the two methods, frequency modulation has found more application than phase modulation.

Wide-deviation F-M is confined to ultra-short waves for the following reasons. Firstly, the frequency band occupied is much greater than the modulating frequency and there will only be room for such transmissions within the ultra-short wave band, with its enormous range of available frequencies and limited range of transmission. Secondly, there would be difficulties in the design of transmitter and receiver to deal with the large frequency band unless the carrier frequency was high. Thirdly, long-distance short-wave transmissions suffer distortion due to selective fading and multiple paths (to be discussed in Chapter V) and this distortion increases with increasing band-width of transmission.

Broadcasting by wide-deviation F-M (such as 75 kc/s) has only been developed extensively in the U.S.A., where it is used to provide short-range broadcast services, of high fidelity, around large cities. The provision of similar stations in this country has been discussed and experimental transmissions are now being made.

F-M is being used to an increasing extent for ultra-short-wave communication circuits, such as those used by police forces for linking up cars with stations. When only " commercial-

quality" speech is required, a deviation of 15–20 kc/s is commonly employed.

On some short-wave telegraph circuits "on-off" keying is employed but the radiation on "mark" has frequency modulation of a tone imposed upon it. In consequence, the "mark" transmission is spread over a band of frequencies and selective fading effects (see Chapter V) are not so serious.

Having thus delineated the possible fields of application for F-M, we can discuss the advantages which may be gained in this limited field.

If we consider first the transmitter, on F-M this has to provide a constant R.F., whilst with A-M the peak power may be four times the mean. It will be clearer when Chapter X has been studied that the efficiency will be greatest when the transmitter is giving its full output, and this can be the case all the time if F-M is employed. Also, a transmitter capable of giving the same peak output, whether A-M or F-M, will produce twice the modulation-frequency power at the receiver if F-M is employed.

Turning now to the receiver, one of the most important considerations will be the ratio of signal to noise at the output. The nature and sources of the components making up the noise output of a receiver are dealt with in Chapter XIV, but it may be stated here that such components are of two classes, those which are continuous and can be resolved into an infinite number of small, sinusoidal components at all frequencies, and those which consist of occasional pulses (lasting only a few micro-seconds) which may be of large amplitude. The latter can, of course, be resolved into a spectrum of radio frequencies, and, in both cases, it will only be those components which lie within the pass band of the receiver that contribute to the noise output.

In an A-M receiver the noise output is produced by R.F. voltages beating with the wanted carrier and thus producing a low-frequency variation of amplitude. The noise has, therefore, the same character as the A-M signal and cannot be separated from it. It is possible to fit a limiter, which reduces the amplitude of large impulses to little more than that of the peak signal.

A type of limiter may also be used which greatly reduces

the sensitivity of the receiver during the short duration of the pulse.

If any other type of modulation is used, then there is a possibility of reducing the noise output by making the receiver insensitive to A-M but still responsive to the type of modulation being received. Thus, in a receiver for F-M, a limiter must be used, and, provided the received field strength is adequate, this can be arranged to "slice off" a large part of the R.F., thus making the receiver unresponsive to amplitude changes. This cuts out the noise voltages during the time that the signal is near its peaks, as well as reducing the value of impulses, as in the case of the A-M limiter.

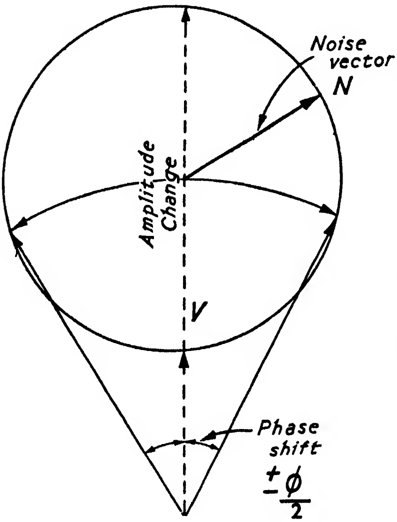


FIG. 46. Illustrating Effect of Noise in F-M Reception.

The noise voltage will, however, be impressed on the F-M in a way which can be seen from the vector diagram of Fig. 46, where V represents a steady carrier and N a sinusoidal component of noise voltage. As in our previous discussion of vectors for modulation, we will allow V to be stationary and N to rotate at the difference frequency. Then we see that, during one cycle of N, the resultant vector swings through the angle ϕ. In the example, N has been made one-half of V and the angle is about one radian. The larger the deviation frequency, the smaller proportion ϕ is of the phase change due to the wanted signal. The amplitude is seen to change by 3 to 1 but the limiter in the receiver will prevent this affecting the output.

It will be seen that the phase change produced by N is the same whatever the frequency of N, and in the case of phase modulation the improvement of signal/noise ratio over amplitude modulation is proportional to the phase deviation.

In the case of frequency modulation, it has been shown that for a given phase change the deviation produced is directly proportional to the modulating frequency. Hence, if the modulation is due to noise having components at all frequencies, those components nearer to the carrier frequency will produce less frequency modulation, that is, less noise in the output. The distribution of noise frequencies in the output is therefore triangular, instead of being rectangular, as in A-M.

If the deviation frequency is increased, the audio output from the wanted signal will be increase but the receiver will have a greater band-width and will therefore pick up more noise voltages. The frequency modulation produced by these will, however, yield frequencies above the audible range and, therefore, will not add to the noise output.

All the above factors can be shown to produce a signal/noise ratio about 30 db higher than with an amplitude-modulated transmitter of the same peak power, *provided* the deviation frequency is about 75 kc/s and the input signal/noise ratio is fairly good.

If the signal/noise ratio at the receiver input is very poor, conditions are rather different. The extra noise voltages introduced by increasing the receiver band-width to accommodate the increased deviation frequency now interact and contribute to the noise output. In such cases a deviation frequency of about 15 kc/s is suitable and may result in a better signal/noise than A-M could give in the same situation.

Because of the "triangular" noise distribution, we can further improve the signal/noise ratio, when using F-M, by "pre-emphasis" at the transmitter and a corresponding "de-emphasis" at the receiver. By pre-emphasis is meant a graded increase in depth of modulation of the higher-frequency modulation components, which are subject to more noise (as already explained), the balance being restored at the receiver.

We have already mentioned that frequency and phase methods have not the limit of carrier amplitude that is imposed by amplitude modulation. With amplitude modulation we cannot, no matter what power is taken, obtain a much greater percentage modulation than about 95% linearly, and work down to a minimum limit of less than about 5%, due to noise considerations and nonlinearity.

With frequency (and phase) modulation there is no theoretical upper limit. We relate the upper limit to an arbitrary 100%, but in general this upper limit could be exceeded with only a gradual distortion appearing due to circuit limitations, there being no abrupt cut-off as there is with amplitude modulation. This, together with the lower noise level that is possible, means that the ratio of maximum to minimum signal amplitudes that can be handled will greatly exceed the figure possible with amplitude modulation, and in consequence a much greater fidelity is possible without artificial compression (and subsequent expansion at the receiver) of the signal.

For reasons which will be more apparent when receivers have been studied, two F-M transmissions on the same carrier frequency do not interfere with each other so much as with A-M. If the wanted carrier is only about twice the unwanted one, the interference at the receiver output is quite small.

Pulse Modulation

In recent years another method of transmitting a signal has come into use. Instead of transmitting a continuous carrier,

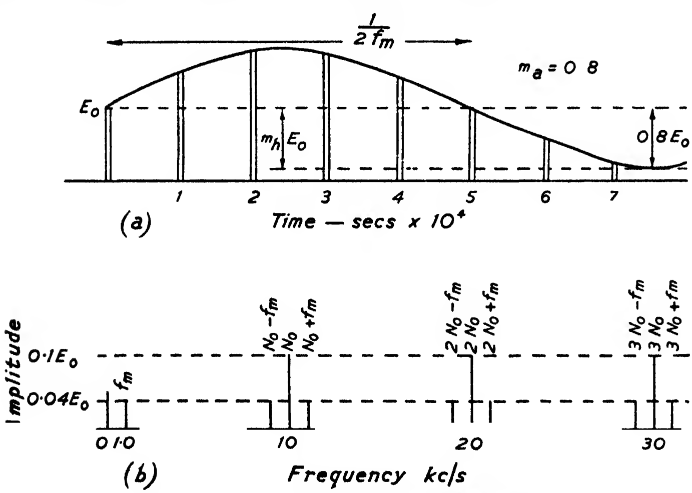


FIG. 47. Pulse-Amplitude Modulation.

short pulses of radio-frequency power are transmitted at regular intervals. These pulses take "samples" or "snapshots" of the signal to be transmitted, and, provided that the

" snapshots " are sufficiently frequent, the receiver is able to reconstruct the signal.

As an illustration, suppose that we are employing a radio frequency of 100 Mc/s and we wish to transmit a sinusoidal signal of 1 kc/s. One way in which this can be done is by transmitting recurrent pulses, each having an amplitude proportional to the instantaneous amplitude of the signal. In our example (Fig. 47a), each of these pulses lasts 5 microseconds and consist of 500 cycles of radio frequency, there being 10,000 pulses per second.

A pulse has other characteristics besides amplitude and these

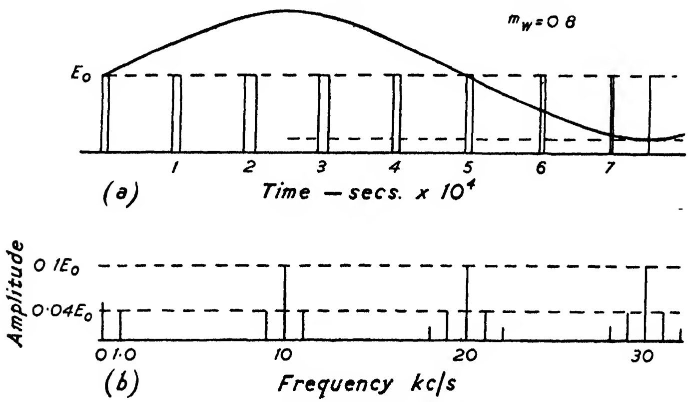


FIG. 48. Pulse-Duration Modulation.

can provide alternative methods of modulation. For example, the pulses can be all of the same height but vary in duration in accordance with the amplitude of the signal. This is known as pulse-duration or pulse-width modulation (Fig. 48a).

The timing of the pulse may be variable, this method being termed either pulse-time, pulse-phase or pulse-delay modulation. Alternatively, the frequency of the pulse may be varied, this method having much the same relationship to pulse-phase modulation as frequency and phase modulation have, when using continuous carrier transmission.

We have discussed the spectrum of a pulse in the previous chapter and have seen that the total frequency band occupied depends upon the narrowness of the pulse and upon the steepness of the growth and decay. There will be a component

frequency at every harmonic of the pulse repetition frequency. Thus, in our illustration, the unmodulated pulse would have components at 10, 20, 30 . . . kc/s and in order to retain an approximately rectangular pulse, these harmonics would have to extend to about 700 kc/s, which is the 7th harmonic of $1/2 \times 10^{-6}$, the frequency equivalent to the pulse duration. Note that it is only the first few harmonics which are of equal amplitude.

If the radio frequency is 100 Mc/s, then the radio transmission would involve frequencies from 99·3 to 100·7 Mc/s.

It will be seen from Fig. 47 that if the pulses are amplitude modulated, there is an f_m component and also each harmonic of N_o is now accompanied by two side bands. The R.F. bandwidth will be the same as before, and, after detection at the receiver, we shall recover the spectrum shown (in part) in Fig. 47b. If a low-pass filter, cutting off all frequencies above the highest modulation frequency it is desired to receive, is placed after the detector, then the desired signal f_m will be obtained, free from other components. The amplitude of the f_m term can be shown to be $E_o \tau_o N_o m_a$.

It will be necessary for N_o to be at least twice the highest frequency component of the modulation. Otherwise the component $N_o - f_m$ will be accepted by the filter and will be a spurious frequency in the output.

When pulse-duration modulation is employed, the relationships are illustrated by Fig. 48. The f_m term and the side bands vary in amplitude in accordance with the signal. As in pulse-amplitude modulation, a low-pass filter will follow the detector and the condition $N_o > 2f_m$ must be fulfilled.

If a modulation index m_w be defined as

$$\frac{\text{Maximum pulse width} - \text{Mean pulse width}}{\text{Mean pulse width}}$$

then the amplitude of the f_m term becomes $E_o m_w \tau_o N_o$.

In our example, the pulses are of $5\mu s$ mean duration and vary from 1 to $9\mu s$ with modulation, so that m_w is 0·8. N_o is 10 kc/s and therefore the amplitude of the f_m term is 4% of the amplitude of the pulse.

Pulse-phase modulation is illustrated by Fig. 49. The modulation index m_p is defined as

$$\frac{\text{Maximum phase displacement}}{\text{Mean interval between successive pulses.}}$$

It can be shown that the amplitude of the modulation-frequency term is $E_0 m_p \omega_m \tau_0$. Recovery of the modulation at the receiver will be more complicated than with the type of modulation

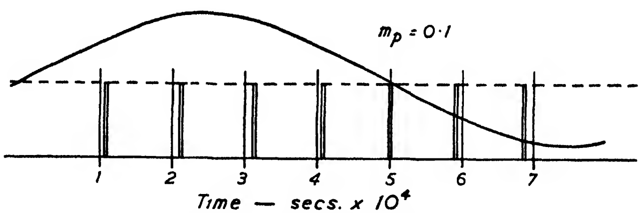


FIG. 49. Pulse-Phase Modulation.

previously discussed because the rise of output with ω_m has to be corrected for.

Let us now consider the very important question of signal/noise ratio when pulse modulation is used. It can be said at once that pulse-amplitude modulation will, in general, result in a poorer ratio than straightforward amplitude modulation,

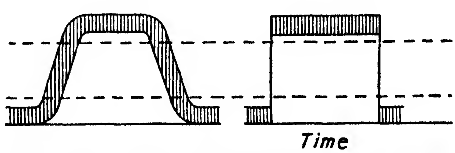


FIG. 50a. Effect of Random Noise on Pulse.

because the receiver will have a wider bandwidth and there are no other compensations.

With pulse-duration or pulse-phase modulation, on the other hand, limiters can be used so that both top and bottom of the pulse is cut off, as in Fig. 50. This results in the loss of some energy, of course. In Fig. 50a the noise is represented as continuous, and it will be seen that noise voltages are only effective during the rise and fall of the pulse. Consequently, the steeper the pulse, the less the noise in the output, as is made clear by considering the ideal rectangular pulse, to which no noise would be introduced.

If pulse-duration modulation is being employed, then it will be seen that noise is effective by increasing the duration of the pulse. On the other hand, if pulse-phase modulation is in use and the leading edge of the pulse is the time reference, then it will be seen that noise alters the phase and gives an output in this way.

Now the steepness of the pulse sides depends upon the bandwidth of the receiver (assuming that this limitation comes before limitations in the transmitter) and, therefore, in general, the wider the bandwidth the less the noise output—the opposite

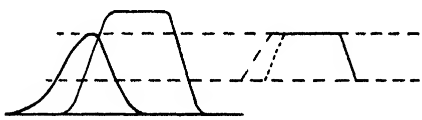


FIG. 50b. Effect of Impulsive Noise on Pulse.

condition to that for amplitude modulation. This is modified by the fact that if we increase the receiver bandwidth we increase the noise *input* but the statement remains true for practical conditions.

If the interfering noise voltages are of an impulsive kind and can be represented as in Fig. 50b, then it will be seen that this pulse may alter either the effective duration or the phase of the signal pulse and so produce output noise.

Systems employing pulse modulation are discussed further in Chapters XIII and XVIII.

CHAPTER IV

THE PROPAGATION OF SHORT AND ULTRA-SHORT WAVES AT SHORT DISTANCES

In this and the following chapter the very involved subject of wireless wave propagation will be treated only in a descriptive fashion. For the sake of completeness, it will be necessary to refer briefly to the longer waves but the main discussion will be concerned with short and ultra-short waves.

It is evidently very desirable that it should be possible to predict the signal-strength which will be obtainable at a distance from a transmitter of specified power and frequency, and using a given aerial system, etc., but the acquisition of sufficient data to do this is one of the most difficult problems in wireless engineering. Theoretical investigations require advanced mathematics and numerical solutions usually depend upon constants which must be determined experimentally. The measurements to determine the constants or to confirm the theories are difficult. Only simplified conditions (as, for example, the assumption of a simplified ionosphere and a smooth spherical earth of uniform conductivity and dielectric constant) are amenable to theoretical treatment, in general.

Wireless telegraphy had been a commercial proposition for some time before any such measurements had been made, but progress in recent years has been rapid, due to the pains-taking research of numerous workers in many countries. As a result of the information collected, a Committee of the C.C.I.R. were able in 1939 to draw up a comprehensive report [14] giving quantitative curves and data covering the propagation, under average conditions, of wireless waves of all wavelengths in use. This certainly does not mean, however, that finality has been reached in the study of wireless propagation, inasmuch as much more information is still required to clear up a number of doubtful points and explain the numerous results which occur periodically and do not conform to type.

In fact the more statistical evidence that is accumulated

the more one realises that the behaviour of the ionosphere is like that of the weather. Whilst one can give guiding rules as to its probable behaviour at a certain place, date, and time, the vagaries of the ionosphere will be certain to produce many incalculable results from time to time.

In the present chapter the propagation of waves along or comparatively near the earth will be discussed, whilst the propagation through the ionosphere will be discussed in the next chapter.

The Hertzian Dipole

The simplest arrangement to consider as a source of radiation is a straight conductor, very short compared with the wavelength produced and carrying a uniform alternating current. The radiation from practical aerials can be calculated by considering the aerial to be composed of a number of such elements.

The equation of the field at a distance from the dipole and on a plane perpendicular to the conductor is given by

$$E = \frac{120\pi I h}{\lambda d} \text{ volts per metre.}$$

where $2h$ is the height of the complete dipole, λ is the wavelength, d is the distance } in metres,
and I is the current in amperes.

It will be evident from the shape of a dipole, that if it is vertical, the radiation will be the same in all directions in a horizontal plane, and this radiation can, therefore, be illustrated by a polar diagram as in Fig. 51a, where the radius represents the field strength which would be measured if a suitable apparatus was carried round in a circle, having the dipole as its centre. A consideration of the field distribution in a zenithal plane shows that there will be no radiation vertically, maximum radiation horizontally, and for other zenithal angles it is reduced, following a cosine law, as shown in Fig. 51b. The distribution in all directions is therefore represented by the three-dimensional figure sketched in Fig. 51c.

It will be observed from the formula that the field strength falls away as the first power of the distance, this reduction of

signal strength simply being due to the spreading of the wave.

At points really close to the dipole, i.e. less than one wave-length, there are also the ordinary electric and magnetic fields to be considered in addition to the radiated fields. As, however, these induction fields are inversely proportioned to the square of the distance, they quickly become negligible and need not be considered in propagation problems generally, but they

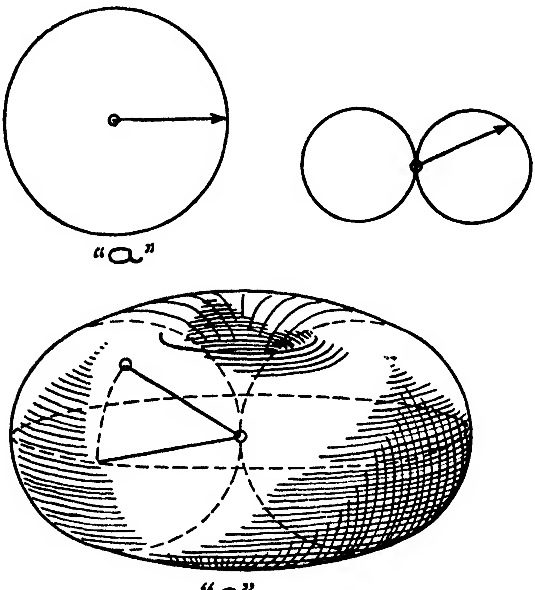


Fig. 51. Polar Diagrams of a Dipole.

are important in problems connected with the behaviour of a number of radiators near to one another. It should be mentioned that the induction fields are in time quadrature with each other, whereas the radiation fields are in time phase.

We now require to know what power is being radiated away from the dipole, and as in the ideal case there is no loss in propagation, we can consider the energy flowing through a sphere at a considerable distance from the dipole. Our equations give the field strength at every point in this sphere and from this we can determine the energy flowing through a

unit area, so that summing over the surface we obtain the total power radiated. The result obtained by this method is:

$$W = \frac{320\,\pi^2\,h^2\,I^2}{\lambda^2}\ \text{watts}$$

Since W is proportional to I^2, other things remaining constant, this power may be considered as used up in a fictitious resistance R_a, such that:

$$W = I^2\,R_a$$

and it will be seen that the radiation resistance of a dipole is given by:

$$R_a = \frac{320\,\pi^2\,h^2}{\lambda^2}$$

The modification necessary to adjust the calculations for the ideal dipole to fit ordinary aerials will be discussed later in Chapter VIII.

The wave at a sufficient distance from such a dipole will be plane polarised, the electric and magnetic fields being mutually perpendicular to each other, and to the direction of propagation. In wireless engineering, the plane of polarisation is stated with respect to the electric field, and with reference to the earth's plane. Thus in a vertically-polarised wave the electric field is in a plane perpendicular to the earth's tangent plane.

If a perfectly-conducting, horizontal, plane sheet is passed through the centre of the vertical dipole, the fields should not be disturbed since the electric field is everywhere perpendicular to the plane. We deduce, therefore, that if we set up a half-dipole on a perfectly conducting plane, the field in the space above the plane is the same as that produced by a complete dipole in space. Another way of arriving at the same result is to suppose that the half-dipole has produced an "image" beneath the plane and that the current in the image is the same in magnitude and phase as that in the actual half-dipole.

Conductivity and Dielectric Constant of the Earth's Surface

In many cases the propagation of wireless waves over the earth's surface is much affected by the values of conductivity (σ) and dielectric constant (κ) and the assumption of perfect conductivity made above would lead to wrong conclusions

being reached. The values of σ and κ naturally vary considerably for different kinds of soil and there are also large variations with weather conditions at the same site. The soil at different depths will also have different properties and it is, therefore, not an easy matter to state an effective value for use in a wave-propagation problem.

Measurements undertaken by the Radio Research Board,[4, 5] showed very large variations in σ with moisture content. Thus for one sample of loam σ was 10^5 electrostatic units ($\rho = 9 \times 10^6$ ohms per cm. cube) when the moisture content was about 1% and $1{\cdot}5 \times 10^8$ (6,000 ohms) when it was 25%, these figures being taken at a frequency of 1,200 kc/s. For the same sample, at the same frequency, κ varied from 3 for a 1% moisture content to 37 for a 25% content.

The conductivity increases with frequency whilst the dielectric constant decreases, the same sample as previously mentioned, when measured at 10 Mc/s, and at a moisture content of 25%, gave values of $\sigma = 2 \times 10^8$ and $\kappa = 30$.

Reflection at the Earth's Surface

We have already seen that if a vertical half-dipole is erected on the surface of a perfectly conducting plane, the field at a distance can be considered as due to a complete dipole. This is a particular case of the "image" theory by which the effect of reflection is taken account of by assuming an image carrying a current $I\ (A\ \underline{|\phi})$, where I is the current in the actual dipole and $(A\ \underline{|\phi})$ depends upon the type of reflection and in this case is $A = 1$, $\phi = 0°$.

The theory of reflection has been completely worked out, so that if σ and κ are known, the reflection coefficient applicable to a wave polarised either in the place of incidence or perpendicular thereto and of any frequency and incident at any angle can be determined. The principal practical difficulty is to know what values of σ and κ to assume, especially if there are changes in the nature of the soil just below the surface. As the significant factors in determining the nature of the reflection are κ and $\frac{\sigma}{f}$ (σ in E.S. units) it follows that the relative effects of κ and σ vary greatly over the range of frequencies used in wireless.

The reflection conditions are, in general, quite different, depending upon whether the electric vector is perpendicular to the plane of incidence (Fig. 52a) or in that plane (Fig. 52b). The former case is simpler and will be dealt with first. If the surface were a perfect conductor then, for all values of θ, $(A\,\underline{|\phi})$ would be -1; that is, the image dipole must be considered as carrying a current, equal in magnitude but opposite in phase, to that in the actual dipole. This is seen to be necessary to satisfy the boundary conditions because there cannot

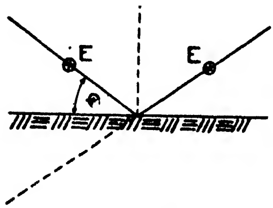


Fig. 52a.

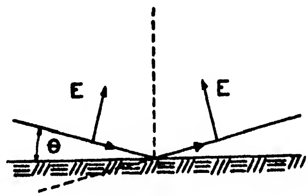


Fig. 52b.

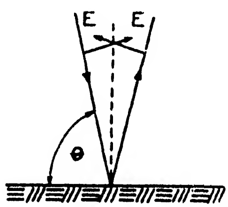


Fig. 52c. Reflection of Plane E.M. Waves.

be any resultant electric field along the surface of a perfect conductor.

If the earth may be regarded as a perfect dielectric ($\sigma = 0$) then there will be a refracted ray passing into the earth, as well as a reflected ray. For small values of θ, $(A\,\underline{|\phi})$ is very nearly -1. For larger values, ϕ is still 180°, but A depends upon the value of κ. Evidently, if κ was 1 there would be no reflection and A would be zero.

If we consider reflection at the surface of sea water ($\kappa = 81$, $\sigma = 10^{-10}$ E.S.U.), then for a frequency of 100 Mc/s $\left(\frac{\sigma}{f} = 100\right)$ the assumption that $(A\,\underline{|\phi})$ is -1 is a very close one for all values of θ and for lower frequencies would be even closer.

For reflection at the surface of earth, assuming $\kappa = 10$, $\sigma = 10^8$ E.S.U., there is a much greater divergence, however, and Fig. 53 (derived from McPetrie's curves [9]) gives values of A and ϕ for $\frac{\sigma}{f} = 1$, and $\frac{\sigma}{f} = 2{\cdot}5$, which correspond (with the values assumed for σ) to 100 Mc/s and 40 Mc/s respectively.

Consider now the conditions for reflection when the electric vector is in the plane of incidence (Fig. 52b and 52c). If the earth were a perfect conductor, then the horizontal components

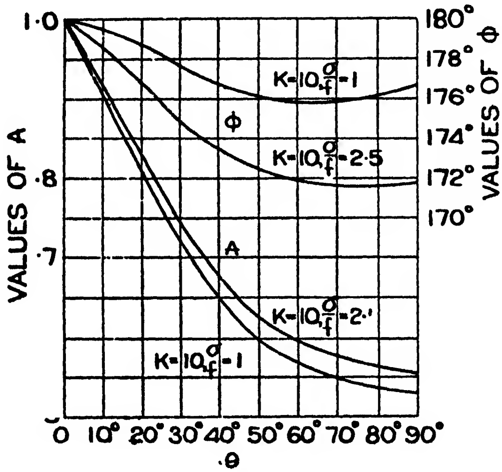


FIG. 53. Polarisation Perpendicular to the Plane of Incidence.

of the electric vectors would have to cancel for all values of θ. This requires that $(A \underline{|\phi})$ would equal $+ 1$. It is evident that when θ is 90° there is no difference between the two polarisations. An examination of Fig. 52c shows that, with the conventional directions indicated, when θ is 90° and ϕ is 0°, the currents in dipole and image will actually be in opposite directions, thus conforming with the statement previously made for polarisation perpendicular to the plane of incidence.

If, on the other hand, the earth may be considered as a pure dielectric, then $(A \underline{|\phi}) = - 1$ for small values of θ (as for polarisation perpendicular to the plane of incidence). As θ increases, however, A at first decreases, becoming zero (that is, there is no reflected ray) at an angle called the Brewster

angle, which depends upon κ, then increases again; ϕ is 180° below the Brewster angle and 0° above.

When actual values of κ and σ are considered, we obtain

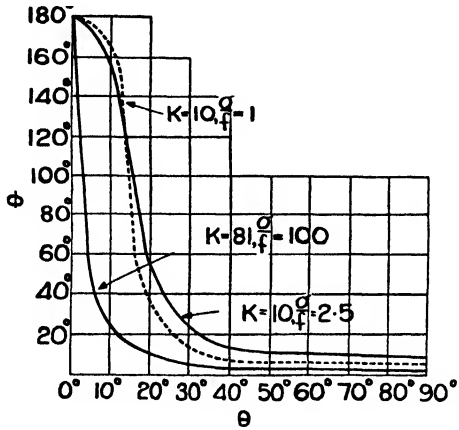


FIG. 54. Polarisation in the Plane of Incidence.

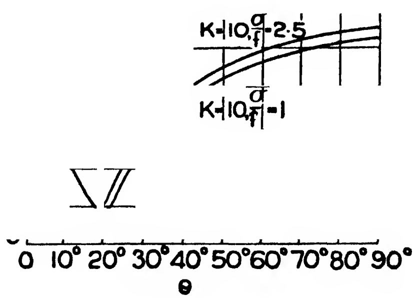


FIG. 55. Polarisation in the Plane of Incidence.

curves such as shown in Figs. 54 and 55, it being observed that with ultra-short waves we approach the condition of $\phi = 0$. From these curves it will be seen that the greater $\frac{\sigma}{f}$ is, relative to κ, the lower is the angle at which A becomes a minimum

and ϕ changes quickly, but in all cases the image and actual dipoles carry equal currents in opposite phase when $\theta = 0°$. It would therefore appear that wave propagation directly along the ground when σ is finite, is not possible, a result which we know to be contrary to experience. This discrepancy will be dealt with later.

The Ray Theory

We can now apply our discussion of reflection at the earth's surface to an approximate theory of propagation which is useful in certain cases.

The state of affairs existing when transmitting over a plane surface from an elevated transmitter T to an elevated

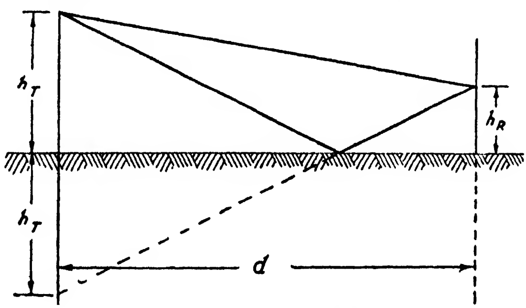


FIG. 56. Direct and Indirect Paths between Transmitter and Receiver.

receiver R may be seen by examination of Fig. 56. The attenuation of the direct ray will be entirely "geometrical," that is, due to its spread, and if T is a dipole then the amplitude of the direct ray will be inversely proportional to the distance.

The distance traversed by the indirect ray is seen to be $\sqrt{d^2 + (h_T + h_R)^2}$ whilst the direct ray travels $\sqrt{d^2 + (h_T - h_R)^2}$ producing a phase difference at R of

$$\frac{2\pi d}{\lambda}\left[\sqrt{1 + \left(\frac{h_T + h_R}{d}\right)^2} - \sqrt{1 + \left(\frac{h_T - h_R}{d}\right)^2}\right] \text{ radians.}$$

If $d \gg h_T$ and h_R, this can be approximated (by the use of the binomial theorem) to

$$\frac{2\pi d}{\lambda}\left[1 + \frac{1}{2}\left(\frac{h_T + h_R}{d}\right)^2 - 1 - \frac{1}{2}\left(\frac{h_T - h_R}{d}\right)^2\right]$$

That is, the phase difference

$$= \frac{2\pi}{\lambda} \cdot \frac{2h_T\, h_R}{d} \text{ radians.}$$

If T' is now given the correct value and phase, in accordance with the curves given, the resultant field at R can be calculated.

If T is a horizontal dipole perpendicular to the plane containing the propagation path, then, whatever the value of θ, the incident wave is polarised perpendicular to the plane of incidence. By using curves in Fig. 53 we can therefore find the resultant field at R and we could plot a polar diagram for the horizontal dipole at T.

Consideration will show that, except for communication with nearby aircraft, or between very high, closely-adjacent hills, θ will not exceed about 10°. It follows from Fig. 53 that we may take $(A \underline{|\phi})$ as equal to -1 in many cases.

If T is a horizontal dipole but its axis is not perpendicular to the plane containing the propagation path, then the incident wave will contain components polarised in the plane of incidence as well as perpendicular thereto and it will be necessary to treat the components separately. The horizontal dipole has directive properties and if remote from earth does not radiate along its length. When near earth a vertically-polarised wave is radiated in this direction.

Suppose now that T is a vertical dipole. For all values of θ this will produce a wave entirely polarised in the plane of incidence, and by reference to curves such as those of Figs. 54 and 55 we can find the resultant field at R. For very small values of θ we see that we can again assume that $(A \underline{|\phi})$ is -1, but it is evident that this approximation is only true over a very small range of θ, and, furthermore, may be invalidated by an effect to be discussed in the next section.

Since the values of κ and σ are so variable and are frequently not known at all closely, it is useful to have roughly approximate polar diagrams for a vertical dipole above the earth's surface. For values of $\frac{\sigma}{f}$ occurring in the short or ultra-short wave band, Fig. 54 suggests that for small values of θ the best simple assumption is to take $(A \underline{|\phi})$ as -1, whilst for higher angles $(A \underline{|\phi})$ should be taken as $+1$. Polar diagrams obtained

in this way (but for a half-wave aerial, not a dipole) are shown in Fig. 150, page 268.

Reverting now to a consideration of Fig. 56 we see that if θ is small and $(A \underline{|\phi})$ may be taken as -1, since the distance travelled by the direct and reflected wave is almost the same, the resultant field strength at the receiver will be much less than that due to the direct ray alone.

The phase angle will be $\pi + \frac{2\pi}{\lambda} \cdot \frac{2h_T h_R}{d}$ and both E_D and E_R components will be of equal strength. A vector diagram will show that the total field E is therefore given by $2E_D \sin \frac{2\pi\, h_T\, h_R}{\lambda d}$.

It will be seen that if $\frac{h_T h_R}{\lambda d}$ becomes $0, \frac{1}{2}, 1 \ldots$ the total field becomes zero, provided that the conditions for which the expression is obtained hold good.

Thus if a receiving aerial at a fixed distance is raised, the received field strength will pass through a series of maxima and minima. Similarly, if a receiving aerial is maintained at a fixed height but brought nearer to the transmitter, the received field strength will fluctuate.

If $\frac{2\pi h_T h_R}{\lambda d}$ is small, we can put the sine equal to the angle (in radians) so that $E = 2E_D \frac{2\pi h_T h_R}{\lambda d}$.

The direction of radiation from T will be practically normal to the aerial and hence, using the expression given on page 91,

$$E = 2\left(\frac{120\pi\, Il}{\lambda d\ .\ 2}\right)\frac{2\pi h_T h_R}{\lambda d}$$

$$= \frac{240\pi^2\, Il\, h_T h_R}{\lambda^2\, d^2},$$

where l is the effective length of the aerial at T.

The above formula must be used with caution and with a clear realisation of the assumptions upon which it rests. When valid this expression will be approximately true whether the dipole is vertical or horizontal, that is whether vertically or horizontally polarised waves are being used, but in the latter

case it will only give the field strength in a direction normal to the dipole.

From the foregoing discussion it will be evident that the ultra-short waves will give good results if signalling is to be carried out between elevated sites such as between steep hills, because the effective values of h_T and h_R will be large.

A consideration of Fig. 57, however, will show that even in such a case there may easily be more than one reflected wave. In the case shown, one reflected ray will largely neutralise the direct ray in the way previously discussed, but this leaves one reflected ray unopposed and therefore the field at R is much greater than indicated by the equation. It is, therefore, possible for the received strength to be large or

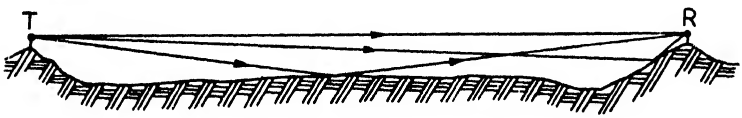


FIG. 57. Multiple Paths between Transmitter and Receiver.

very small and, when setting up a receiving site, the aerial should be moved about to find the best position.

It will also be evident that a small change in the transmission frequency will alter the length of the different paths, as measured in wavelengths, and will alter the phase-angles between the various components at the receiver and give a different resultant field-strength. If the transmitter is radiating a modulated wave, therefore, it does not follow that the various frequencies will preserve their correct relationship to the carrier, either in magnitude or phase. This effect is likely to be pronounced when very short waves are used for television.

Multiple Path Effects in Television

We have already seen that multiple paths may produce rapid variations of signal strength with position or with frequency and that the frequency response curve of a television receiver may be seriously distorted thereby.

Since the television modulation frequencies are so high, it is also possible for the difference in time of propagation over different paths to be comparable with the reciprocal of the

modulation frequency. This effect has been studied theoretically by Lawson.[16]

Energy arriving by a second path will evidently give a completely separated double image, or "ghost," if the delay in transmission equals the time between two picture elements. With present British standards this will be about one-tenth of a micro-second, corresponding to a difference in path length of 30 m.

In the case of telephone transmission, an "echo" delayed by so small an amount behind the main signal would, of course, merge into it and be unnoticed.

It is unlikely that the earth reflections previously discussed would produce "ghosts" unless h_T or h_R are unusually large

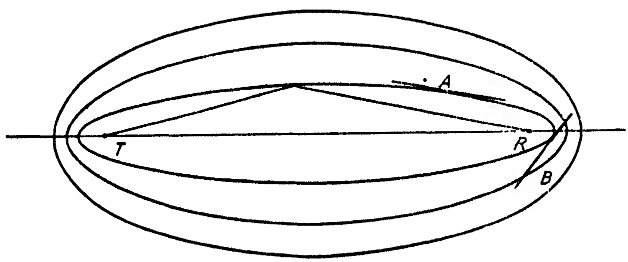


FIG. 58. Echoes from Multiple Paths.

and d small but reflecting surfaces, such as buildings, may well do so. Reflecting objects capable of giving the same difference in path lengths will lie on an ellipse having transmitter and receiver at the foci and a set of conical ellipses can therefore be drawn for different delays.

Clearly-defined double images will be given by buildings such as A (Fig. 58) which lie approximately along one of the ellipses, whilst an object which lies across ellipses, such as B, will produce a blurring effect on the screen of the television receiver.

In practice the most disturbing position for a reflecting object will be on the flank or behind the receiver, because it will then reflect a powerful signal into it. The transmitting aerial will normally be high and reflections from buildings near it should be unimportant. An improvement can, therefore, be produced by using a directional aerial at the receiver.

The effects become worse if the radio frequency is increased

because smaller objects become effective radiators, but it is easier to employ directional reception. An increase in definition means, of course, that shorter path differences are effective in producing "ghosts."

The Surface Wave

All that has been discussed hitherto concerning reflection and propagation assumes that, at the point where reflection occurs, the incident wave is a plane wave. If the dipole is sufficiently near to the earth (as measured in wavelengths), however, the incidental wave has not become plane before the surface is reached, and under these conditions the "image" theory will not give the total field.

In such cases there is a surface wave, the foot of which travels along the earth's surface. This wave undergoes attenuation because it produces currents in the earth, and there will be a wave travelling through the earth.

The attenuation depends in a complex way upon κ and $\frac{\sigma}{f}$. For the lower frequencies it is the conductivity that is important, the higher this is the lower being the attenuation. For frequencies within the short and ultra-short wave bands κ is also an important factor, the higher the value of κ the lower the attenuation. Thus for all frequencies the surface wave is less attenuated over sea than over land. The attenuation increases with frequency and becomes very great in the ultra-short wave band.

The magnitude of the surface wave from a vertical dipole is such that the ray theory previously mentioned ceases to be a useful approximation unless the dipole is about one wavelength above earth and it can be seen, therefore, that it is only likely to be of use for ultra-short waves.

When a horizontal dipole is used, however (and we are concerned with propagation in a direction perpendicular to the axis of the dipole), the surface wave is of much smaller magnitude and the optical theory with the image in anti-phase is a useful approximation even when the dipole is only $\frac{\lambda}{4}$ above earth.

When a wave is travelling over an imperfectly conducting

earth, the electric field must have a component along the direction of travel because the induced currents in the earth require potential differences over the surface to produce them. There will also be a wave travelling through the earth and this absorbs energy. The electric field of a vertically polarised wave is therefore no longer exactly vertical but is tilted, the foot lagging behind. The angle of tilt is a function of σ and κ and by measuring the tilt σ and κ can be deduced.

Another way of considering the matter is to realise that since there is a loss of energy at the foot of the wave there must be a downward supply of energy from the upper part of the wave, to balance matters up.

The surface wave is able to travel beyond the optical range, that is, to follow the curvature of the earth, because of two effects: diffraction, and refraction in the lower air layers. It is well known that light rays " leak " around an opaque object and illuminate an area that should be in shadow if a strictly straight-line propagation only was considered. The magnitude of this diffraction effect increases with the wave length and it will be evident that diffraction will be an important feature of wireless wave propagation, since the waves are immensely longer than light waves.

Any theory allowing for diffraction must assume, of course, that the earth is a smooth sphere and hills or other obstructions may greatly influence the signal strength actually obtained. Such local obstacles will have more effect on the shorter wavelengths because diffraction is less marked, and the nearer the obstruction is to either transmitting or receiving aerial the more influence it will have.

The original diffraction theories were worked out for transmitting and receiving aerials on the earth's surface. Elevation to any appreciable fraction of a wavelength is clearly impracticable on the longer waves and would not be markedly beneficial.

As the wavelength is reduced, however, the range of a transmitter on the earth's surface to a receiver also on the earth's surface becomes exceedingly small and hence, when ultra-short waves are being used, either transmitter or receiver (or both) will usually be raised to a height which may be several wavelengths.

It is evident that the "ray" theory cannot be employed for ranges greater than the optical because diffraction is not allowed for.

Eckersley's Theory of Ultra-short Wave Propagation[10]

T. L. Eckersley has extended the diffraction theory to deal with the most general case of transmission from a raised vertical

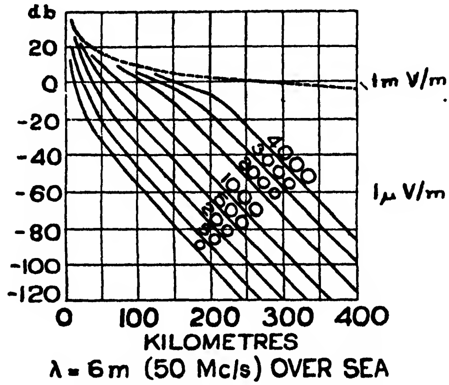


FIG. 59. Propagation of Ultra-Short Waves over Sea.

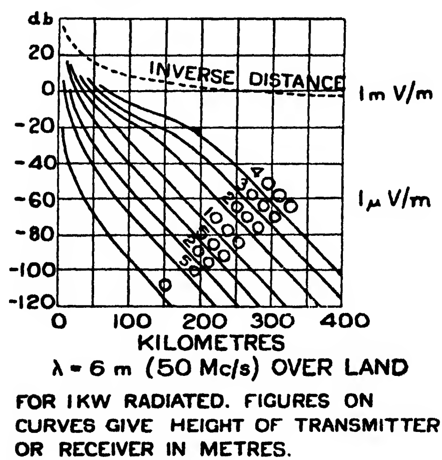


FIG. 60. Propagation of Ultra-Short Waves over Land.

dipole, around the curvature of the earth (allowing for the earth's σ and κ) to a raised receiving dipole. The results of the theoretical analysis have been presented in a set of curves

giving the field strength to be expected at different distances from a transmitting dipole on the earth's surface, for different heights of receiving aerial. Alternatively, the curves will give the field strengths at the earth's surface from a transmitting dipole raised to different heights. These curves, of which two sets are prepared, one for transmission over sea, and one for transmission over land, are shown on page 704, but Figs. 59 and 60 give one set for 6 metres, and it will be seen that the transmission of such waves over the sea (Fig. 59) is distinctly better than over land (Fig. 60), and that the field

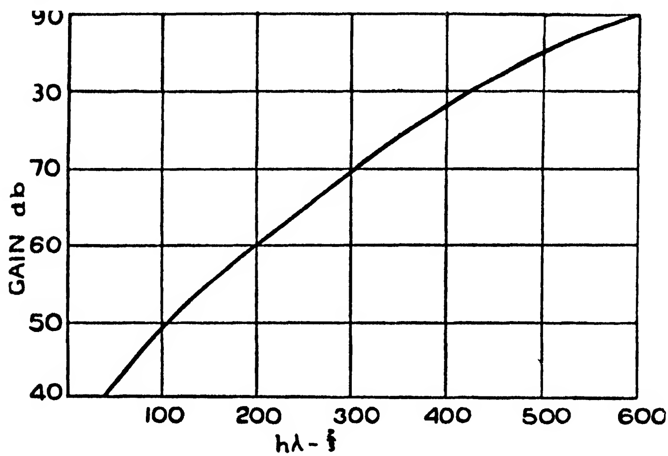


Fig. 61.

strength obtained increases rapidly with height of transmitting (or receiving) aerial.

The use of the ray theory in cases where it is not even approximately true, has led to the belief that signals will get suddenly weaker when we pass beyond the optical range. The full theory shows no such sudden change, except that at great heights (where the ray theory is more applicable) the curves do bend over and show increased attenuation after the optical range is exceeded.

Eckersley has produced an auxiliary curve giving the gain with height, so that a value of field strength can be obtained when both transmitter and receiver are raised. He finds that above a certain height, h_0, such that $h_0 \lambda^{-\frac{2}{3}} = 50$, the gain

with height is practically independent of the earth's constants σ and κ and that it is therefore possible to give a curve of gain against $h\lambda^{-\frac{2}{3}}$ as Fig. 61. The gain below h_0 is, however, very dependent upon σ and κ and in order to find the total gain due to raising to a height h, it is necessary to subtract a constant number of db from the given curve, in accordance with the table below:

Wavelength (Metres)		2	4	6	8	10
Db to be subtracted	Over Land	0	2·2	3·7	4·9	5·9
	Over Sea	13·6	18·0	20·8	22·9	24·4

An example may make the use of the curves more clear:

Required to calculate the field strength at a distance of 50 km from a 50 watt transmitter employing an aerial (assumed to have a cosine polar diagram) at a height of 50 m. The receiving aerial is on an airplane at a height of 1,000 metres. The transmission is over land (assumed $\sigma = 10^{-13}$ e.m.u., $\kappa = 5$) and the wavelength is 6 metres.

From Fig. 60 the received energy on the ground at 50 km from a 1 kW transmitter at 50 metres height is 44 db below the datum (i.e. the energy corresponding to 1 mV/m). The actual power being only 1/20th of that assumed in the curves will reduce the received energy by 10 log 20 = 13 db, thus making the F.S. -57 db.

For this case $h\lambda^{-\frac{2}{3}} = 303$ and from Fig. 61 the gain due to the height of the airplane is 71 db, from which 3·7 db have to be subtracted (see Table), and hence the energy becomes

$$-57 + 71 - 3{\cdot}7 = 10{\cdot}3 \text{ db above datum.}$$

If the field strength in mV/m is x, then:

$$+\,10{\cdot}3 = 20 \log x \text{ from which}$$

$$x = 3{\cdot}3 \text{ mV/m.}$$

In Appendix I will be found a series of curves for wavelengths between 2 and 6 metres.

Propagation through the Troposphere

The lower part of the earth's atmosphere is called the troposphere because within it the atmosphere is supposed to be well mixed by winds and to be of the same chemical constitution. The density and temperature then decrease uniformly with height, the temperature drop being about 1° C. per 100 m.

When an E.M. wave passes from one medium to another. the refractive index,

$$\mu = \frac{\sin \phi_1}{\sin \phi_2} \quad \frac{v_1}{v_2} = \sqrt{\frac{\kappa_2}{\kappa_1}}$$

if the permeability of both media is the same. The refractive index of a gas (referred to a vacuum) is given by $\mu = A\delta + 1$ if μ is not far removed from unity. A is a constant and δ is the density of the gas. The connection between pressure, ρ, density, δ, and absolute temperature, T, for a perfect gas is $\rho = R\ \delta\ T$ (where R is the gas constant).

In the case of the atmosphere, we can work out from the temperature gradient the variation of δ and hence μ. As μ decreases with height, it follows that an E.M. wave travelling through the troposphere will follow a curved path, which for

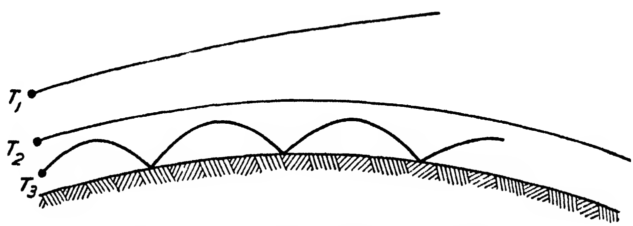


FIG. 62. Abnormal Refraction Effects.

" normal " conditions has a radius about five times that of the earth.

Refraction in the troposphere therefore assists diffraction over the surface of the earth, to make it possible for waves to travel beyond the horizon.

Because the dielectric constant of water is so high (approximately 80) the value of κ is very dependent upon the water vapour present. The specific humidity (that is, grammes of water vapour per kilogramme of air) should be independent of height in a well-mixed atmosphere.

Conditions in the troposphere are often very different from the " normal " conditions which have been assumed. Particularly at times of calm, " good " weather, the air tends to stratify and it is found that the temperature does not fall at the usual rate for the first 1,000 metres or so. Also, the humidity decreases rapidly with height. Under these condi-

tions, κ decreases much more rapidly with height than for the "normal" conditions and the curvature of the path may be greater than that of the earth's surface.

The height over which κ is varying abnormally is sometimes only about 30 m. and is seldom more than 1,000 m. Hence, if we consider transmitters at various heights (see Fig. 62) it will be seen that rays from a transmitter near the earth's surface may be trapped and travel far beyond the horizon by successive reflections. The phenomenon is known as super-refraction. It is as if the wave has travelled within a "duct."[12] If the aerial is placed at a greater height the ray will be lost, as the bending is less.

It is found that abnormal ranges of this kind are common on centimetre wavelengths, occasional on waves of a few metres and are never obtained on medium wavelengths. If we treat the propagation from the "ray" point of view, the reason for this does not appear. If we consider it more from the wave point of view, however, we can find an explanation.

The "zig-zag" propagation within the "duct" is much like that in a wave guide (see Chapter VII). For the wave guide there is a limiting frequency below which transmission is not possible, and the same is true for the duct. Owing to the entirely different boundary conditions in the two cases, whereas a wave guide can transmit waves for which the guide is only half a wavelength wide, the duct in the atmosphere has to be large compared with the wavelength.

Hence ducts are often formed which guide centimetre waves efficiently. Much deeper ducts are formed occasionally which are able to guide waves of a few metres but never ones which are capable of retaining medium waves. The theory of these ducts is due to Booker.[12]

Whenever refraction is making an important contribution to the signal strength at the receiver, we shall expect the signals to vary in strength—that is, to "fade." Slow fading will occur as the ducts come and go with changing temperature, etc., and there will also frequently be rapid fading, due to sudden local changes in the atmosphere. In cases where there is more than one path between transmitter and receiver, we shall expect rapid fading as atmospheric changes cause differences in path lengths.

Absorption in the Atmosphere

Since light waves are so greatly absorbed by fog, and to a less extent by rain, it is natural to enquire how short radio waves must be before they are appreciably affected.

An average fog consists of very small droplets of about 0·01 mm. diameter and spaced about 4 mm. It is a matter of common observation that raindrops vary greatly in size but an average drop during moderate rain has a diameter of about 1 mm. and drops are roughly 100 mm. apart.

We might expect, therefore, that neither rain or fog would affect radio waves to a measurable extent unless the wavelength is below 10 cm and that absorption due to rain would be greater than that due to fog. Light waves, on the other hand, are so much shorter than the diameter of even fog particles (wavelength of green light, $0{\cdot}5 \times 10^{-3}$ mm) that fog is more effective than rain, because of the greater concentration.

Some time before the production of sufficient power on centimetre wavelengths to make their use practicable, Stratton examined theoretically the magnitude of absorption to be expected. More recently the problem has been worked out by Ryde.[11]

Both agree that absorption will be small for 10 cm waves, even with tropical rain or dense fog. Attenuation increases rapidly for shorter wavelengths and tropical rain would produce 2 or 3 db/km attenuation of a 3 cm wave, increasing to about 15 db/km at 1 cm. With the rainfall usual in this country, however, the values would only be about a tenth. A fog which reduces visibility to 30 m should introduce an attenuation of 0·23 db/km at 3 cm and 2 db/km at 1 cm, which is about the same as that due to a moderate rainfall.

Some experimental results are given on page 118.

Summary of Factors influencing Propagation

The theoretical work on short and ultra-short wave transmission (excluding transmission through the ionosphere) may be summarised as follows :

The range of a transmitter on the earth's surface to a receiver, also on the ground, becomes smaller as the wavelength is reduced. Under certain conditions theoretical curves are available from which the field strength at an elevated receiver,

produced from an elevated transmitter, may be deduced. The earth's surface has to be assumed smooth and spherical and likely average values of σ and κ assumed.

Refraction in the troposphere may produce an appreciable increase in signal strength at longer distances and can be expected to produce fading. Very long ranges may be expected at times on very short waves, due to unusual variations of dielectric constant with height.

An approximate " ray theory," though of limited application, is useful and simple in certain cases and shows us that where communication takes place over uneven ground or obstructions commensurate with the wavelength, there are frequently several paths by which energy can reach the receiver and in consequence an interference pattern can be produced.

When wavelengths such as 3 cm are employed, rain and fog may be expected to introduce a noticeable attentuation.

We now propose to study very briefly some of the more important experimental investigations which have been published, in order to see how far the theories outlined are correct and complete.

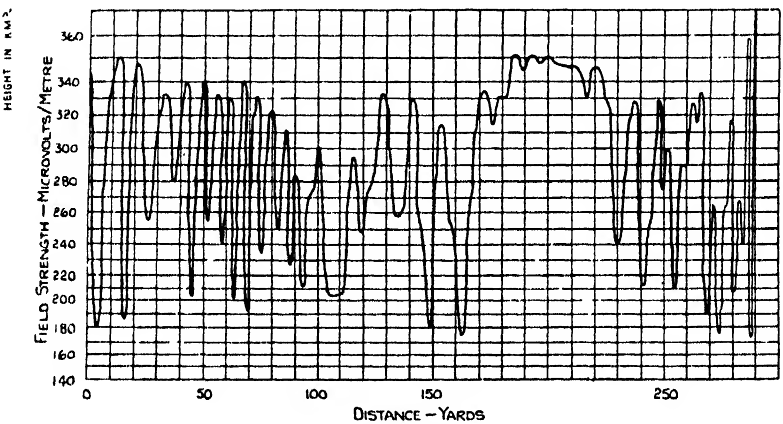
Experimental Studies of Transmissions of 5 to 7 Metre Waves

Very many investigations have been made on these wavelengths either to check the theories discussed or to provide data for the establishment of a commercial service.

All investigators have found that signal strength varies rapidly with receiver position as the " ray " theory suggests. In endeavouring to check theoretical field-strength/distance curves, therefore, it is necessary to choose as open a site as possible and it may be necessary to take several readings at sites very near together and take a mean. Any site chosen must be well away from all power or telegraph lines, railways and trees, and should preferably be uniform in character and flat.

Many measurements [18] have been made on transmissions at several frequencies in the neighbourhood of 50 Mc/s (6 metres) from the Empire State Building, New York, the aerials of which are the highest fixed aerials in the world (400 metres above earth) and the results obtained and published by Jones are in good agreement with the Eckersley diffraction curves.

Jones[18] and other workers have observed the large fluctuation in signal strength which can be produced by moving cars, etc. At these wavelengths it is evident that such objects are sufficiently large in comparison with the wavelength to re-radiate to a considerable extent. It has also been observed that aircraft can produce noticeable fluctuations of signal strength, even when at some distance from the receiver. Since the aircraft is in a much stronger field than the receiver (due to its height) re-radiation from it is considerable. As an



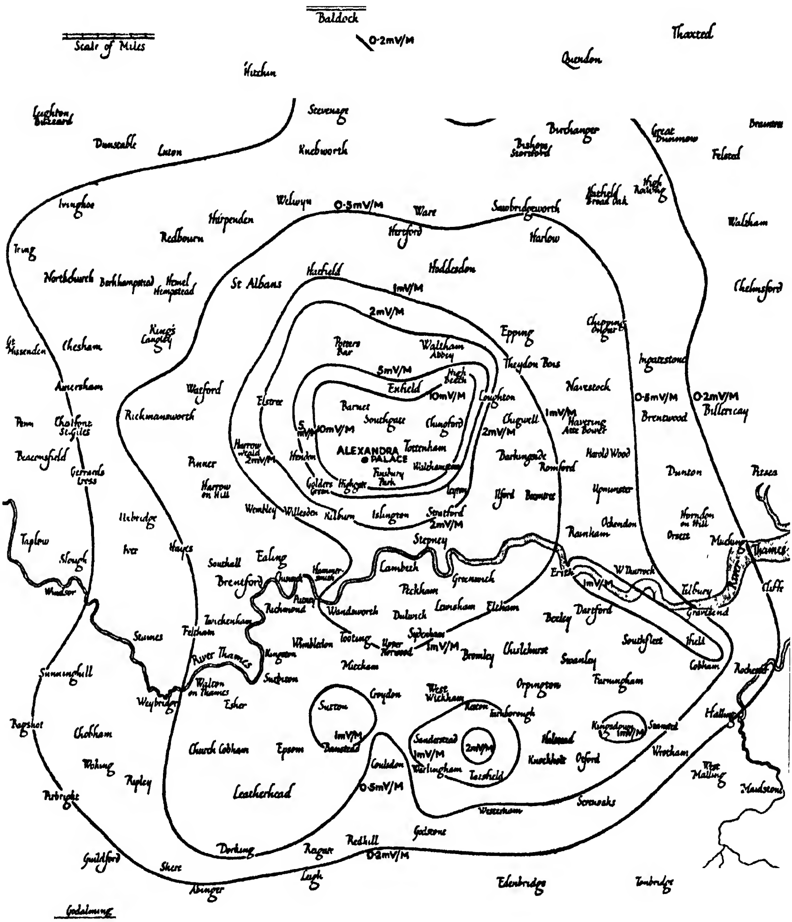
* Fig. 63.

* By permission of the British Broadcasting Corporation.

instance, television reception at Chelmsford (30 miles from the transmitter) often suffers rapid fading due to aircraft between Chelmsford and London, the picture brightness varying as the changing position of the aircraft changes the phase of the re-radiated wave relative to the direct wave. Tests in rural districts also show rapid variations of signal strength with position, especially near trees. A typical curve obtained by the B.B.C. (Fig. 63) shows the large variations obtained for small changes of position.

Maclean[19] studied 50 Mc/s transmissions at three sites, one of which was within the "optical" range, one 700 feet below the line of sight, and one 11,400 feet below. He found that even within the optical range there was slow fading up to 10 db in amplitude, whilst at the longer distances fading was much

more pronounced, the maximum values of the field strength being about 20 db above the average value. There was no correlation between the fading at the three sites. The average



* FIG. 64.

* By permission of the British Broadcasting Corporation.

field strength was much higher at night, probably due to refraction.

It has been demonstrated by George [20] that the frequency response curve of a wide-band receiver (such as for television) can be greatly modified by multiple-reflection effects.

An extensive study[15] of the transmissions on 6·67 metres (45 Mc/s) (vision) and 7·25 metres (41·5 Mc/s) (sound) from the B.B.C. Television Station at Alexandra Palace, London, has been made and the B.B.C. have published "contour" maps, one of which is shown in Fig. 64. From these maps, curves for field strength in the different directions, N. S. E. and W., have been plotted in Fig. 65, and compared with Eckersley's theoretical curve, showing the close agreement.

The British Post Office have made a number of investigations.

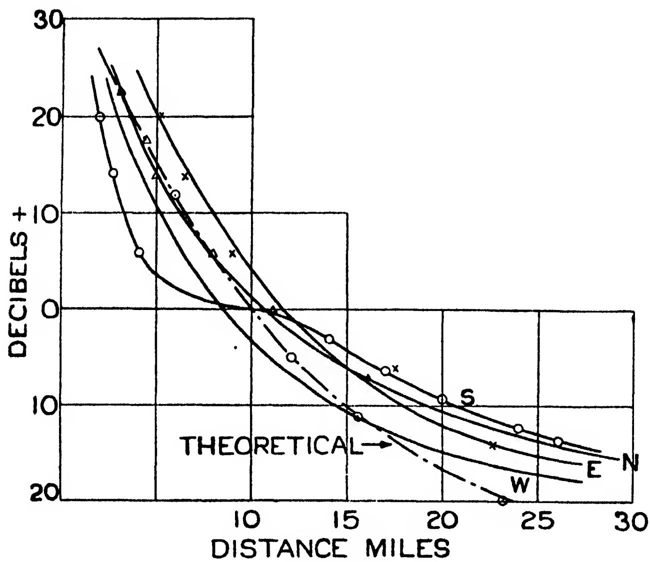


Fig. 65.

A continuous, two-year record of signal strength over a 137 km path (mainly over sea) has been analysed by Smith-Rose and Strickland, the wavelengths concerned being 5 and 8 metres.

It was found that fading was slight during periods of low barometric pressure and was more severe in "good" weather, which is in accordance with theory concerning refraction. It was difficult to decide whether the fading was entirely due to changes in refraction over a single path or whether it was partly due to interference between a direct ray and other rays which were reflected from discontinuities in the troposphere.

The gain of signal strength with height above the earth

has been measured by a number of observers, and in particular the work of Jones [18] using an auto-gyro aircraft have been shown to agree closely with Eckersley's theoretical curves.

Experimental Studies of 1 to 3 Metre Waves

The ray theory in the form given by McPetrie and Saxton [7] has been checked by them, using wavelengths of 2 and 3 metres, and very good agreement has been obtained. Conditions were such that the ray theory, with image in anti-phase, was a reasonable approximation. It was confirmed that on an open site, vertically and horizontally polarised waves gave the same field strength and that signal strength was directly proportional to receiver height as stated in the theory. It was observed that the 2-metre transmissions were attenuated to a greater extent when passing over London, but the 3-metre transmission appeared to be unaffected.

Continuously during one year, Burrows [21] and his associates studied the propagation of a 2-mbtre (150 Mc/s) wave over a "non-optical" 60 km path and found fading up to 20 db The average field strength was higher during the night but the fading more pronounced. Conditions were very similar at a distance of 200 km.

Wavelengths within this range have been used extensively for radar and the possibility of exceptional ranges under certain weather conditions amply demonstrated. As far as is known, the world's record for long-range radar is held by a 1½ metre station at Bombay which frequently "sees" ships at 1,000 km distance and the coast of Arabia at 2,000 km, during the hot season, but reverts to its normal range during the monsoon.

Experimental Studies of Waves Shorter than 1 Metre

Some of the pioneer work on these wavelengths was done by Yagi and Uda [22] in 1928 and in 1931 the I.T. and T. Corp.[23] commenced to study the propagation of 18-cm waves across the Straits of Dover. In these tests between fixed stations, well within the optical path, signals were steady when weather conditions were steady and were unaffected by rain, fog or snow. Abrupt changes of temperature and pressure often produced fading up to 40 db, particularly on still, summer

days. There was almost certainly a reflected ray from the water and fading was partly due to changes in relative path length. There was some evidence that the rise and fall of the tide affected the signal by altering the length of the reflected ray path.

Marconi and his assistants [25] conducted an extensive investigation of 50-cm waves in 1932. On one typical test over sea, good telephone signals were received at 93 km (the optical range being 84 km) but very deep fading, allowing only of occasional reception, was then experienced up to 140 km. At this distance signals improved considerably and remained good up to 160 km, being finally lost at 175 km. During a later test, consistent reception was possible at distances up to five times the optical range even when hills intervened, and signals were reported at nine times the optical range.

Somewhat similar results were obtained by Hershberger [24] in 1934. Using 75 cm, he obtained useful telegraph signals at 140 km over sea, this being five times the optical range, as the shore transmitter and the receiver on shipboard were low. Up to 30 km the signals were strong and steady, from 30 to 50 km became weak and fading and above 50 km the strength varied little with distance and there was some fading.

These long ranges, using such short waves, attracted considerable attention at the time. The "duct" theory, developed during the war, seems an adequate explanation. It seems likely that the poorer results at the intermediate distances, noted by both investigators, were really due to meteorological changes whilst the tests were in progress.

A great deal of data has been obtained by the operation of radar on 10 and 3 cm bands. In addition, a number of stations were set up in this country, expressly for studying propagation.[26] One of these networks, which is still in operation, comprises two transmitters on the Welsh coast, near to each other, but *A* is at 165 metres above sea level, whilst *B* is only 27 metres. At a distance of 92 km, a pair of receiving stations were set up, *C* being at 252 metres and *D* at 29 metres. At a distance of 320 km (in S.W. Scotland), in nearly the same direction from *A* and *B* as are *C* and *D*, a further pair of receiving stations were installed. *E* is at a height of 115 metres and *F* at 29 metres.

At both A and B there are 9 and 3 cm transmitters, giving power outputs of 0·6 and 0·15 watts respectively. Paraboloid reflectors of 120 cm diameter are used at the stations. The signal strength measurements are more accurate than has previously been possible. Measurements of atmospheric pressure, temperature and humidity are made by aircraft flying over the route at varying heights.

The reader will need to consult the original papers for details of the results obtained but the following general conclusions emerge. Field strengths at C from A (0·89 optical range) agree fairly well with theoretical values, assuming a "standard" variation of refractive index with height and taking account of reflection from the sea and of diffraction, though fading is present. The same is true of the signals over the paths BC and AD, which are 1·21 and 1·40 times the optical range respectively.

The signals at D from B (2·4 optical range) were, however, nearly always considerably greater than calculated for a "standard" atmosphere. Here the path will lie near the earth's surface and appears to traverse a duct which remains fairly constant. The average level was about 35 db below that for the path AC and the values obtained over a day may vary by 50 db. The greater usefulness of elevated sites when reliable communication is required is clear.

Over the 320 km path, the "standard" level of received signal at E and F would be far below the minimum value which the receiver could cope with (in the case of the B to F transmission on 3 cm, about 500 db below) but signals were received at times.

During three typical summer days, measurable 9 cm signals were received at E from A (3·82 optical range) for about 70% of the time and these sometimes exceeded the "free-space" value. This is worked out for the transmitter and receiver considered as isolated in space but allowing a likely value for the attenuation in the atmosphere. Over the same period, the 3 cm signals followed the 9 cm ones in general but were weaker, presumably due to greater absorption.

During this typicals ummer period the 9 cm transmissions from B were received at F (8·45 optical range) for 50% of the time and sometimes approached the free-space value. The

3 cm transmissions were only received for two brief periods but then approached the free-space value.

The results confirm the general impression that signals beyond the optical range tend to be larger during still, summer weather because ducts are more easily formed. Especially-low field strengths often follow the formation of fog but this is probably not due to increased attenuation but because the formation of water droplets has reduced the invisible water vapour in the air and hence the dielectric constant.

Other tests which have been made include transmission from a low site to the top of Snowdon, to see whether the formation of ducts at low levels would result in a large drop of field strength at a height, but this did not appear to be the case. Transmissions have also been made across London on 9 cm and the signal strength found to be about the same as for a similar path in the country.

The absorption of 3 cm, 1 cm and 0·6 cm waves by rain has been investigated by Robertson, King[27] and Mueller.[28] It is not easy to get consistent results as, if a lengthy path is chosen in order to make the attenuation appreciable, then the rainfall will seldom be uniform. On the other hand, if the path is short it is necessary to measure small increases of attenuation accurately.

The results are summarised in the table below.

TABLE VII. *Absorption due to Rain*

	Rainfall.				
Wavelength.	Light 1 mm./hour.	Moderate 4 mm./hour.	Heavy 15 mm/hour.	Cloudburst 100 mm./hour.	
3·2 cm. 1·09 cm. 0·6 cm.	? < 0·6 < 0·6	< 0·3 0·6 to 1·2 1 to 2	< 0·6 3 to 6 4 to 6	2·5 to 3 20 to 30 25 to 35	Attenuation (db. per km.)

It will be seen that absorption of 3 cm waves by rain will not often be serious. Moderate rain can produce an appreciable effect at 1 cm, however. The increase in attenuation for wavelengths less than 1 cm is not so great, which is in accord with theory.

Application of Short and Ultra-short Waves over Short Distances

The available band of short wavelengths is so valuable for long-distance communication that the use of such waves for short distances must be limited. In tropical countries, however, long and medium waves are so seriously interfered with by atmospherics that short waves are sometimes employed for broadcasting or comparatively short distance communication. It is proposed to transfer such services to ultra-short waves.

Short waves have been employed in the past for such services as police radio, but, here again, ultra-short waves are now usually employed. Difficulties caused by the great variability of signal strength with change of receiver position have been largely overcome. Frequency modulation is often used.

Ultra-short waves are very useful for point-to-point communication over distances up to, say, 100 miles, especially if transmitter and receiver can be on an elevated site. The radiation can be readily confined into a narrow beam, interference is small and the equipment compact. Such services are quite satisfactory at distances considerably in excess of the optical range if the receivers have automatic gain control devices.

Waves below one metre were used for some important channels of communication during the war and are likely to come into increasing use now that reliable transmitting and receiving apparatus is available.

Selected references are given at the end of Chapter V.

CHAPTER V

THE PROPAGATION OF SHORT AND ULTRA-SHORT WAVES THROUGH THE IONOSPHERE

WHEN, in 1901, Marconi successfully demonstrated that communication across the Atlantic (actually between Poldhu, Cornwall, and Newfoundland) could be achieved with electro-magnetic waves, not only did he place the first milestone on the road to long distance wireless communication, but he opened up a new field of thought regarding our earth's surrounding atmosphere and stimulated scientific interest in a direction which has continued up to the present day.

For this experiment a wavelength of about 1,300 metres was used and the theories outlined in the previous chapter are insufficient to explain the long range obtained. In fact the theoretical figure for the field strength at a great distance on a perfectly conducting spherical earth was only 10^{-8} of that probably obtained.

Assuming the earth to be a spherical conductor, Heaviside (and Kennelly) conceived an upper ionised layer forming a second spherical conductor concentric with earth, between which is a homogeneous insulating medium. These two spaced conductors thus form a spherical transmission line, so that any electrical disturbance across the line creates plane polarised electro-magnetic waves which are propagated through the insulating medium between these boundaries, and such waves, therefore, automatically follow the earth's curvature.

For a number of years there was no very direct experimental evidence for the existence of this ionised layer except that crude measurements of field strength showed that this was of the order predicted by approximate theory. From 1925 onwards, however, many workers have studied the problem by various methods. As a result, it has been found that there are, in fact, several effective "layers" of ionised air and the whole region has been termed the ionosphere. In this

book only a simple treatment of the subject will be attempted, dealing particularly with practical application to wireless communication, and the reader is referred to the very extensive literature (references to which are given at the end of this Chapter) for more detailed information.

Polarisation of E.M. Waves

We have already noticed that a vertical dipole set up on the earth's surface produces a vertically polarised wave, that is, one in which the electric field is vertical. The magnetic field is horizontal, so that the direction of propagation, the electrical field and the magnetic field, are mutually perpendicular.

After the passage of a wave through the ionosphere it may be found that the plane of polarisation is revolving. If the value of the electric field in this plane remains constant, it may be resolved into two equal components at right angles to each other and in time quadrature. The magnetic field consists of two similar components and the wave is said to be *circularly polarised*.

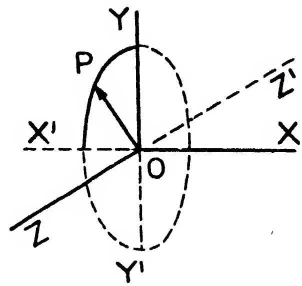


FIG. 66. Vector Representation of Polarised Wave.

If the electric field is represented by a vector, say, OP, Fig. 66, then if the wave is circularly polarised, this vector must be regarded as rotating at a uniform rate as it progresses forward. Thus its extremity traces out a spiral, and the projection of this spiral on a plane normal to the direction of propagation will be a circle as shown. Its components along OY and OX are therefore given by $OP \sin \omega t$ and $\pm OP \cos \omega t$, respectively, the sign of the second component depending upon the direction of rotation of OP.

The amplitude of OP may vary as it revolves, so that its extremity traces out a spiral whose projection is not a circle but an ellipse, and the wave is then said to be elliptically polarised, circular polarisation being merely a particular case of this. As before, this resultant can be resolved into two components along OY and OX, but these two components will no longer be of equal amplitudes.

The Passage of an E.M. Wave through the Ionosphere

In deriving the relationships which follow, the rationalised M.K.S. system of units will be used.

Suppose an E.M. wave having an electric field $F \sin \omega t$ to be traversing ionised air in which there are N free electrons per m^3. Let Q be the charge on an electron and m its mass. The electric field will exert a force $FQ \sin \omega t$ on the electron and, if v is the velocity produced, then

$$m \frac{dv}{dt} = FQ \sin \omega t \text{ and } v = -\frac{FQ}{\omega m} \cos \omega t \text{ m/sec.}$$

The motion of N electrons per m^3 constitutes a current density NQv or

$$-NQ \frac{FQ}{\omega m} \cos \omega t \text{ amps./m}^2.$$

In a vacuum the displacement current produced by $F \sin \omega t$ would be

$$\kappa_0 F\omega \cos \omega t$$

and the total current density is, therefore,

$$\omega F \cos \omega t \left[\kappa_0 - \frac{NQ^2}{\omega^2 m} \right] \text{amps./m}^2.$$

The ionised air is behaving as a medium of dielectric constant

$$\kappa = \kappa_0 - \frac{NQ^2}{\omega^2 m}.$$

The refractive index of the ionised air (relative to un-ionised air) will be given by

$$\mu = \sqrt{\frac{\kappa_0 - \dfrac{NQ^2}{\omega^2 m}}{\kappa_0}} = \sqrt{1 - \frac{NQ^2}{\omega^2 m \kappa_0}}$$

If we substitute values for Q, m and κ_0

then $$\mu = \sqrt{1 - 80{\cdot}5 \frac{N}{f^2}}$$

A wave entering the ionised air will therefore be refracted away from the normal to the surface as shown in Fig. 67.

If N is increasing with height, then the wave will travel in a curve, and if the bending is sufficient, the wave will be directed back to the earth again.

We cannot trace out accurately the path of a wave through

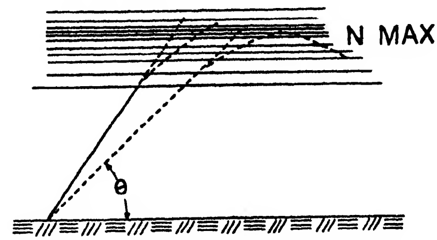


FIG. 67. Refraction in the Ionosphere.

the ionosphere because we do not know sufficiently well how N varies with height.

The velocity of an E.M. wave is given by $\frac{1}{\sqrt{\mu\kappa}}$ (where μ here is permeability) and if c is the velocity in un-ionised air, the phase velocity in the ionised air becomes $\frac{c}{\sqrt{1 - 80{\cdot}5 \frac{N}{f^2}}}$ and the group velocity becomes correspondingly lower than c, being given by c^2/v_{ph}. It is to be noted that the velocity now depends upon f.

For a brief discussion of phase and group velocity, see page 243.

As we make our wave enter the ionosphere more steeply it evidently requires to be bent through a larger angle if it is to be returned again. No rays will be returned above a certain angle given by $\cos\theta_o = \sqrt{1 - 80{\cdot}5 \frac{N_{max}}{f^2}}$ though, if the frequency is low enough, θ_0 may be 90°. As the frequency is raised we eventually reach a value for which no wave is returned from the ionosphere, however oblique the incidence.

A consideration of the geometry will show that, due to the earth's curvature, even a ray which leaves the transmitter tangentially to the earth's surface cannot enter the ionosphere at less than a certain value of θ, which depends upon the height

of the lower edge of the ionosphere. It will be seen, also, that if a ray travels in a straight line it will make a continually increasing angle with the tangent to the earth below. The smallest angle which a ray can make with the E layer is about 8° and with the F layer about 14° (see page 131).

To gain some idea of numerical values we may note that if N is 4×10^{11} electrons per m^3 and f is 20 Mc/s, then $\mu = 0{\cdot}96$, phase velocity $1{\cdot}04\ c$ and group velocity $0{\cdot}96\ c$.

If N_{max} is 4×10^{11}, then the critical frequency for vertical incidence will be 5·7 Mc/s, whilst the highest frequency which would be returned for $\theta = 10°$ would be 33 Mc/s (9·1 m). These values are approximate, due to the simplification of theory.

In the above discussion we have ignored certain factors and it now becomes necessary to discuss these. It will be evident that positive ions in an ionised gas will also experience a force on the passage of a wave. Their mass being so very much greater than that of the electrons, however, their motion is small and their contribution to the total current negligible.

The Effect of Collisions in an Ionised Region

When the electrons are set in motion by the wave, collisions will be caused with the gas molecules and the motion of the electrons will be modified. For the highest frequency the average time between collisions is so long compared with the period of the wave that the effect of collisions is not so very important. In our simple analysis the electron motion was such that an entirely quadrature component of current was added to the displacement current and hence κ remained a simple quantity and there was no absorption of energy in the ionosphere, any more than in un-ionised air. When the effect of electron collisions is allowed for, however, κ becomes a complex quantity. The total current in the ionosphere has now a component in phase with the electric field, namely a conduction current, and this draws energy from the wave, which is therefore attenuated. The calculations to obtain the effective dielectric constant are now much more complicated, but we can gain an idea of the result by considering the average effect of many collisions as providing a damping force proportional to the motion of the electron but opposing its motion. We have something akin to the mechanical vibration of a body

having inertia and friction under the influence of a simple harmonic force. When there are a number of cycles of the wave during the average period between collisions, the frictional force is small and the electron motion is nearly in quadrature with the applied electric force as we have already seen, that is to say the ionosphere is acting as a pure dielectric.

If, however, the frequency is made much lower, the frictional force is large and the electron motion becomes nearly in phase with the electric field, that is the conduction current predominates and at such frequencies the ionosphere behaves as a good conductor rather than a dielectric. In this case an incident wave is therefore reflected and only a small amount of energy, which is quickly dissipated, enters the layer.

Thus the average collision time of the layer marks a dividing line between two types of wave propagation. This time is such that the medium broadcast frequencies lie near the dividing line and thus will be expected to have characteristics common to both long and short waves.

The Effect of the Earth's Magnetic Field

A second fact that has been ignored is that the ionosphere is situated in the earth's magnetic field. In consequence, if an electron is set in motion by the electric field of a wave, it may have another force acting on it due to the earth's magnetic field. If we consider a plane polarised wave entering normally a uniformly-ionised medium, two special cases arise according as the direction of propagation is along or transverse to the direction of the magnetic field.

In the former case, shown in Fig. 68, the electron motions become elliptical instead of " to and fro " as previously assumed, and it can be shown that, as a result, two circularly polarised waves are produced which travel with different velocities through the ionosphere and suffer different attenuations. In the case when the direction of propagation is transverse to the earth's magnetic field, two particular cases are shown in Figs. 68a and 68b.

In 68a is the electric vector in the incident wave is along the direction of the field. The motion of the electron is therefore along the magnetic field and is therefore unaffected by it, and as though the earth's earth were absent.

In Fig. 68c the electric vector is transverse to the earth's field, and the motion of the electrons becomes elliptical in the plane at right angles to the field. The wave remains plane polarised, but with a velocity modified by the presence of the earth's field. If the electric vector is neither transverse nor along the earth's field the wave is resolved into two plane polarised waves, one of the type in 68a and the other of the type in 68c.

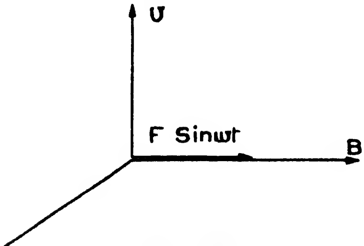


Fig. 68a.

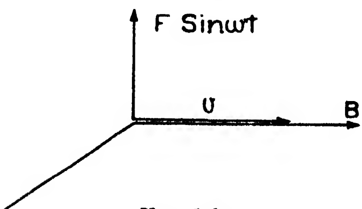


Fig. 68b.

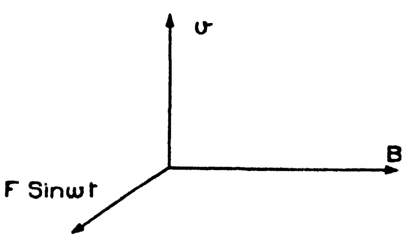


Fig. 68c. Illustrating the effect of the Earth's Magnetic Field.

In the general case when the direction of propagation is oblique to the direction of the earth's field the wave is resolved in the medium into two elliptically polarised waves (whose characteristics depend also upon the density of medium), and in the case of wireless transmission through an ionosphere of varying electron density, in which the direction of the wave is altering with respect to the direction of the field as the wave proceeds, the polarisation characteristic will change from point to point along the ray path. The polarisation of the wave on emergence will therefore depend upon the angle the ray makes with the earth's field as it leaves the ionosphere.

As the frequency of the wave is varied the elliptic paths of the electrons vary in shape because, for the same value of F the forces on the electrons due to F and B change in magnitude. At a certain frequency termed the resonance frequency, the elliptic orbits become very large and hence the number of

collisions is very great, and the losses are greater than at other frequencies. For a value for B of $0{\cdot}5 \times 10^{-4}$ webers/m^2 the resonance frequency is 1·40 Mc/s (214 metres).

At lower frequencies the magnetic field reduces the amplitude of the electron motions, thus reducing the collisions and hence the attenuation. At higher frequencies the main effects are those already discussed and at the shorter waves, especially for the very oblique incidence such as obtains with long-distance transmission, the earth's field has little effect.

Measuring the Properties of the Ionosphere

Existing methods of investigating the ionosphere are only capable of measuring the heights of the maximum electron densities. Before 1939, the properties had only been investigated consistently at a few places on the earth, the most intensive work having been done by the Bureau of Standards, Washington, U.S.A., the Radio Research Board at various places in England and by the Marconi Company, Chelmsford, England. Studies had also been made in India, Australia, the Polar Regions, and at Huancyo, Peru. The last named place was chosen because it is on the earth's magnetic equator, and this simplifies in some ways the interpretation of the results obtained.

During the war, in order to increase the reliability of communications, a considerable number of new ionospheric observations were set up all over the world and our knowledge of the ionosphere greatly increased thereby.

The existence of the ionosphere was indirectly shown by the fact that the various series of long-wave signal-strength measurements agreed reasonably well with theories dependent upon a spherical conducting " ceiling " around the earth, but the first direct measurements from which the effective height could be estimated were made by Appleton and Barnett in 1925 on about 300 metres. A transmitter was set up whose frequency was continually varied, and the received signal strength at a point 88 km away was found to vary through a number of maxima and minima as the phase difference between the surface ray and the downcoming ray varied. When investigations of layer height were first commenced, the existence of an F layer was not suspected, but Appleton found a sudden

discontinuity in measured height at certain times and deduced the existence of an upper layer.

Breit and Tuve in America first developed a method of measuring layer height which was adopted by Appleton and has since been used by most workers on the subject. A very short " pulse " signal is transmitted and is received at a station a mile or two away. The surface wave is therefore received and also a ray which leaves the transmitter almost vertically, a small portion of which may be " reflected " from the ionosphere. By the use of an oscillograph at the receiver, the time elapsing between the arrival of the surface wave and

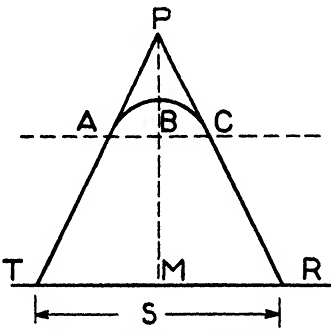


FIG. 69. Equivalent Height of N max.

the reflected ray can be accurately determined and their relative amplitudes compared.

It is necessary to understand clearly the meaning of the height of a layer deduced from such measurements. Whilst the reflected ray is in the layer its group velocity will be reduced below c, whilst the velocity of the surface wave will equal c throughout its journey. If an equivalent path be assumed for the reflected ray which it is supposed to traverse with velocity c, then if the length of this path be l and the surface distance s, we have the relation $\frac{l-s}{c} = t$ where t is the interval between the reception of the two signals. It can be shown that the equivalent path is the " optical path " $TAPCR$ (Fig. 69), so that the equivalent height measured is MP.

As the frequency on which the pulses are transmitted is made greater, we shall reach the critical frequency for vertical

incidence. When θ_o is 90°, the expression on page 123 for cos θ_o shows that $N_{max} = 1{\cdot}24 \times 10^{10} \times f_c^2$, where f_c is the critical frequency in Mc/s.

Thus exploring the ionosphere with a variable frequency, it is possible to obtain the critical frequencies of the various layers as they become transparent at any time and place to waves of vertical incidence, and Fig. 70 shows a typical critical-frequency height curve taken at Chelmsford at midday, in the summer.

It will be noticed from Fig. 70 that the critical-frequency

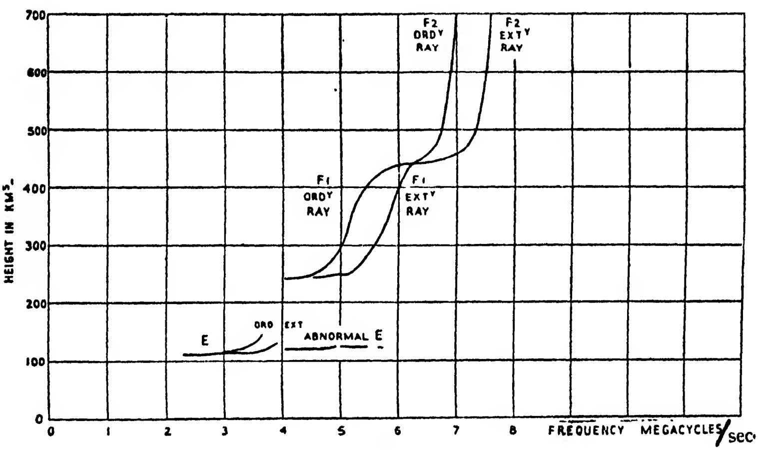


FIG. 70. Height of Layers in Ionosphere.

height curves split into two branches as the critical frequency is approached. This is due to the "echo" being split into two signals separated from each other by the action of the earth's magnetic field. They are known as the ordinary and extraordinary rays, and the separation of the critical frequencies is dependent upon the strength of the earth's magnetic field. In the northern hemisphere, at vertical incidence, the ordinary ray emerges from the layer with a left-hand sense of polarisation when observed from above and the extraordinary ray with a right-hand sense, whereas in the southern hemisphere the reverse is the case.

It is evident that in wireless communication we are usually interested in the behaviour of waves incident obliquely on the

ionosphere, whereas the pulse measurements deal with vertical incidence. The pulse measurements have also to be made at lower frequencies than many of those used for long distance communication.

A very approximate relation between the two is seen from page 123, but methods for more accurate prediction of the behaviour of obliquely-incident rays from the data given by pulse measurements have been developed by Martyn, N. Smith, and Millington.

The first observations of oblique rays were those obtained by T. L. Eckersley, using the facsimile-telegraph apparatus. By its use he was able to measure the time interval between different rays arriving at a receiver after having been reflected a different number of times from the earth (see page 157). From these measurements the maximum electron density in the layer concerned may be deduced.

The Structure of the Ionosphere

Having considered the methods by which the ionosphere has been "explored" we now turn to consider what these methods have revealed.

The ionisation of the atmosphere is brought about by ultra-violet light and corpuscular radiations from the sun. Hence, the number of free electrons present will increase in any part of the ionosphere which is exposed to sunlight. During hours of darkness, recombination will be taking place, that is, free electrons will be uniting with positive ions to form neutral molecules, and at low atmospheric pressure this recombination process is slow and hence the density of free electrons will decrease continuously during the hours of darkness.

The density of free electrons at any time and place varies with height above the earth's surface and experimental evidence suggests the presence of several layers. Or more precisely, the number of free electrons per m^3 rises to more than one maxima with height as shown pictorially in Fig. 70, where frequency in Mc/s is proportional to electron density. These layers are enumerated E, F, etc., the notation being introduced by Appleton, who first discovered the F layer.

There are two principal layers in the ionosphere, namely

E and F, the latter, however, often separating into two parts, the F_1 and F_2 layers respectively.

The lower, or E layer, remains constant in height, at about 100 km, but its density varies with the sun's altitude. Thus it will be most dense at midday in summer, and of minimum density during the night. Similarly the F_1 layer height remains constant at about 200 km, although it is not always observable as a separate layer owing to its merging with the F_2 layer at certain times; but, as far as can be judged, its density, like

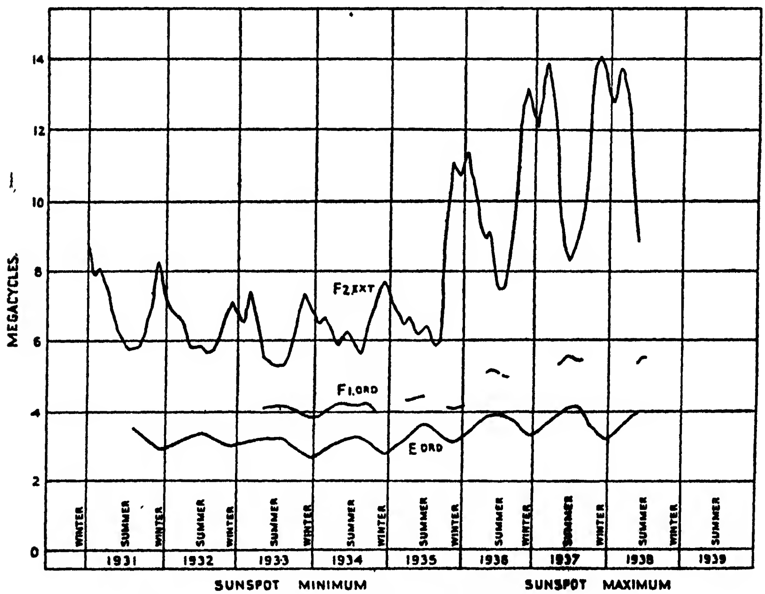


FIG. 71. Variation of Electron Density with Time.

that of the E layer, can also be directly correlated with the sun's altitude. It is assumed that both the E and the F_1 layers are formed by ultra-violet radiation from the sun, as both these layers are absent during the winter periods at the polar regions.

The F_2 layer does not appear to be caused entirely by ultra-violet radiation as it can be observed in the polar regions during the winter months, and its general behaviour is abnormal. First of all, its height is greatest in the summer daytime, and it descends to the level of the F_1 layer in winter. Secondly,

the density of the F_2 layer, in contrast with the E and F_1 layers, is found to be lower in the summer months than during the winter and it also suffers from an annual as well as a bi-annual effect. Thus it is found that the height of the F_2 layer is decreased and its general density increased in *both* northern and southern hemispheres, between September and March. These inconsistencies of the F_2 layer are seen in Fig. 71, which correlates the equivalent layer density with years.

It is sometimes found that the density of the E layer increases greatly for a short time. These sporadic conditions are local in character and the agent producing them is not definitely known.

Besides the principal layers mentioned, there is evidence of layers lower than the E layer, known as the C and D regions. The existence of a G layer beyond the F_2 layer has been suggested as an explanation of some echoes which show a very great equivalent height, but other more probable explanations have also been advanced. There is some evidence for a D region at between 50 and 90 km., and a C region between 20 and 35 km., both these layers having very variable characteristics. Certain workers have obtained pulse measurements which indicate the permanent existence of a series of thin layers of nearly constant height and electron density, extending from 1 km. upwards to about 15 km. The existence of these layers has been deduced entirely from pulse measurements, but at these heights direct observation of the electrical state of the atmosphere can be made by means of balloons, and observations do not reveal such ionised regions. There are also theoretical considerations which make it difficult to accept the existence of layers having the characteristics suggested, and the interpretation of these pulse results is, therefore, a matter of doubt.

The actual values of electron density in the various layers depends not only upon diurnal and seasonal changes, but upon the eleven-year solar cycle, which also affects the earth's magnetic field. During the years when sun-spots are most numerous, the electron density becomes greater.

The ionosphere has been investigated over sufficient places on the earth to show that the condition of the E and F_1 layers is much the same all over the earth when sunlight

conditions are the same but, as already mentioned, this is not true of the F_2 layer.

The results of continuous pulse-measurements made at Washington are summarised each month in the P.I.R.E. and a prediction made of the maximum usuable frequencies for the succeeding month.

We are now in a position to correlate the general facts of propagation theory for communication over great distances. We have shown that the three main layers of the ionosphere, namely E, F_1, and F_2, vary very considerably with time and season, and their condition will determine which has the maximum control on any frequency used for long distance communication. Before showing how the many variable factors which control communication may be co-ordinated into workable charts, or characteristics, to enable us to predict the correct wavelength to use for any given distance at any given time, we will briefly summarise the general features of long distance communication on the different wavelengths.

The gamut of the present useful wireless spectrum may be considered to extend from 30,000 metres down to a fraction of a metre, and as has been mentioned in Chapter I this spectrum can arbitrarily be divided thus :—

(1) The long and medium wave band	=	30,000–400 metres.
(2) The critical wave band	=	400–150 ,,
(3) The short wave band	=	150–10 ,,
(4) The ultra short wave band	=	below 10 metres.

The Long Wave Band

Since both the earth and the E layer behave as good conductors to long and medium waves, their propagation is substantially represented as a " spherical transmission line " type. The wavelengths being commensurate with the height of the lower layer, the state of vertical electric polarisation with which long waves are usually emitted persists (nearly) throughout their journey. Because the lower layer responds quickly to the sun's action, the day and night effects on long waves follow closely the actual coming of day or night.

Attenuation is the only factor which comes into the consideration of long-wave propagation, and this will be proportional to $f^{\frac{1}{2}}$, hence the possible distance of communication

with a given power is proportional to the square root of the wavelength used. With any given wavelength the signal strength distance curve is of exponential form as shown in

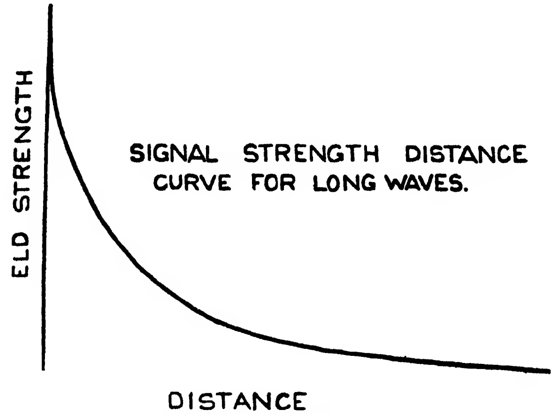


Fig. 72. F.S. Distance Curve.

Fig. 72, and the curve connecting wavelength and communication distance for a given radiated power will be as shown in

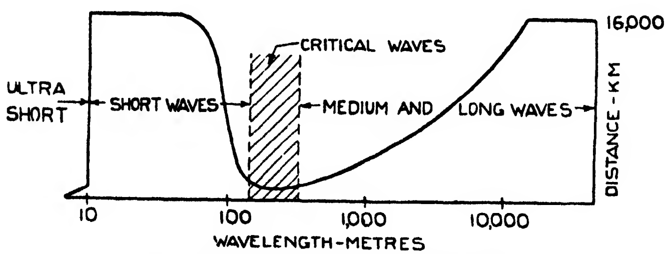


Fig. 73. Wavelength—Range Curve.

Fig. 73. From the second curve it can be seen that only waves above some 15,000 metres are suitable for communication over the greatest distances on earth, namely, half the circumference.

The Critical Wave Band

The critical wave-band includes waves having frequencies near the resonance frequency discussed on page 126. As can be imagined, the transition from long-wave propagation pheno-

mena to short waves is not sudden, but goes through a phase whose condition shows certain characteristics of both types. Waves coming in this class are not confined therefore to frequencies approximating to the collision frequency, but include waves lying in the region between 150 and 400 metres where reliable, long-distance communication is impossible.

During the daytime communication is confined to an area covered by the surface ray, which has a short range, and waves in this band are therefore suitable for short-range services, such as broadcasting, where the intention is to serve a limited area. During the dark hours, however, there is considerable reflection from the upper layer and hence signals may be transmitted over a considerably increased area. This longer, night range is not necessarily useful because at points where the surface and reflected waves are of about the same amplitude considerable fading and distortion results, due to interference between them.

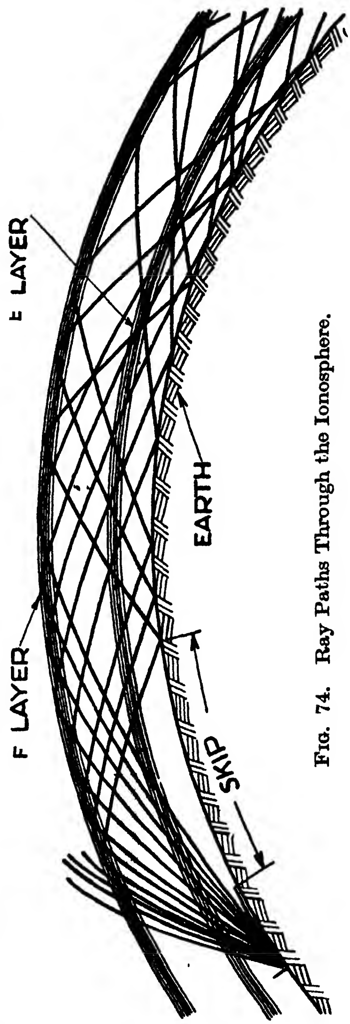


Fig. 74. Ray Paths Through the Ionosphere.

The Short Wave Band

Since short wavelengths are small compared with the height of the ionosphere, it is permissible to use the idea of rays as we do when discussing many problems in optics. It is necessary to remember, however, that the rays which we draw in explanatory diagrams have no separate existence, and are merely representative of an indefinite number of other possible paths. Thus in Fig. 74 each line should be considered as representing

a sheaf or pencil of contiguous rays assumed to spread as the ray is propagated outwards, and if the medium through which the ray pencil passes is homogeneous, the "illumination" of any section of the path is uniform. Short waves are able to penetrate the E layer because there is time for several oscillations of an electron at these frequencies before a collision with a gas molecule takes place, but this is the portion of the path in which the greatest attenuation occurs, because there are more air molecules per c.c. present at the lower altitude and hence more collisions. The rays then travel on to become bent at the F layers, where they also suffer some attenuation. If the frequency is below the critical frequency of the layer at the time and at the oblique angle at which the ray strikes the layer, the ray returns to earth, again passing through the first layer, but if the bending is not sufficiently great the ray escapes through the layer and is finally lost to the earth. The plane of polarisation with which the ray is emitted may not persist and in fact rarely does. Short waves generally will penetrate into the F_1 and F_2 layers, and since the electron densities of these layers do not have any simple relationship with the sun's altitude, variations in short wave conditions, whilst largely dependent on daylight and darkness, do not follow immediately and the connection is therefore less obvious.

The Skip Distance

Energy radiated horizontally from a transmitting aerial near the earth's surface is quickly absorbed due to the large earth losses, as has already been explained on page 105, and hence only short-distance communication can be carried out by this horizontal radiation, which is usually known as the surface ray. Energy radiated at high angles may not be bent sufficiently at the layers to return to earth at all and therefore escapes. Shallow angle radiation at an angle just great enough to escape absorption by the earth will, as shown in Fig. 74, enter the lower layer, suffer attenuation there, be bent at the upper layer and return to earth. A consideration of the figure indicates that between the distance at which the surface ray becomes negligible, and the distance at which the first ray returns to earth from the layer, there may be a zone which is not covered by any rays. This is called the skip zone and the distance

across it the "skip distance." Signals are not necessarily entirely absent in this area, as it will usually be "illuminated" by scattered radiation, to be defined later. Thus the signal

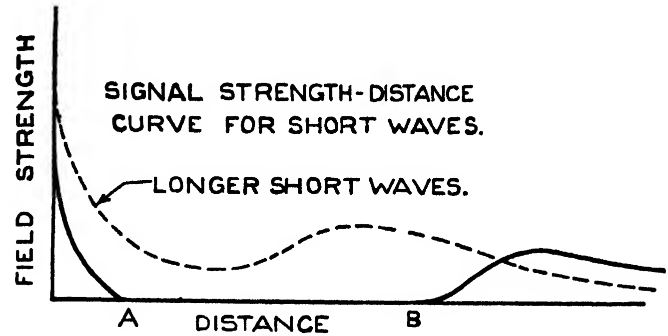


FIG. 75. Field Strength—Distance Curve.

strength-distance curve for short wave communication is of the form shown in Fig. 75, where A marks the limit of the surface

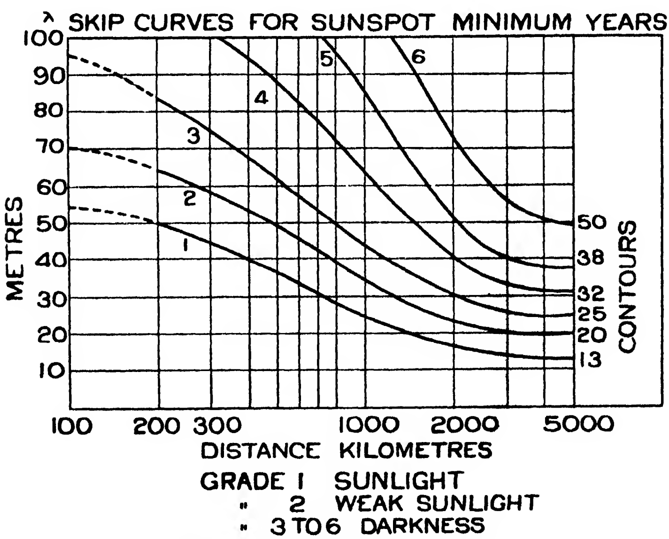


FIG. 76. Skip-Distance Curves.

ray, B the distance where the first down-coming ray is received and AB the skip distance, although it is more usual to consider the distance from transmitter to B as the skip, because the

range of the surface ray is always small. It should be mentioned that unusually short skip distances are occasionally recorded, due to an abnormal condition of *E* layer causing reflections from it.

In the case of the longer short waves (that is in the region of 100 metres) there may not be any definite skip distance but only a pronounced dip in the signal strength-distance curve, as shown in Fig. 75. With wavelengths in the neighbourhood of 12 metres, however, the skip distance may reach several thousand km when most of the path is in daylight, whilst if most of the path is in darkness no ray would return to the earth's surface at all. Thus the skip area will vary considerably in extent depending upon wavelength, season, and time of day, and Fig. 76 shows the skip distance for different wavelengths during daylight, twilight, and darkness conditions. The various grades, 1 to 6, are referred to later on in the chapter. These curves are really concerned with bending, and show the minimum distance at which the bending of a ray is sufficient to return it to earth, and it follows that for a given distance and wavelength the darker the grade the less the bending.

It is to be noted that during the sunspot maximum years the skip curves are lowered by some 25% so that the limiting wavelength for sunlight conditions is reduced to some 9 metres.

Scattering

Within the *E* layer there are local centres or " clouds " of high electron density. Waves incident upon these may be scattered in all directions and hence may provide a very uncertain signal at a receiver. Scattering is more in evidence in the skip area because the only energy received may be due to this cause, but it is nearly always the means by which some of the energy reaches a receiver.

If the transmission is non-directional then scattering will take place from many points situated all round the transmitter. It follows that a direction-finder placed where the energy is mainly received by scattering will not indicate a bearing, since energy is reaching it from many directions.

It has been shown by T. L. Eckersley that in some cases where directional transmission is taking place, a direction finder

placed within the skip zone will indicate the bearing of the station as being more or less where the "beam" enters the layer and not the true position of the station, thus showing a marked scattering of rays from this point.

Exactly how the earth is "illuminated" beyond the skip distance by rays bent back from the upper ionised layer has been a matter for a number of theories from time to time, but the generally accepted one is that the rays continue in a series of ricochets, that is by successive reflections from the earth and refractions from the upper layer, as shown pictorially in Fig. 74. If this is so it will be seen that the earth beyond the skip zone will be illuminated by a whole series of rays, from very shallow-angle rays nearly tangential to the earth's surface which have proceeded by a few ricochets, to rays near the critical angle which have "hopped" many times. The high-angle rays which make many ricochets will, in general, be more attenuated in traversing a given distance over the earth's surface than the low angle rays, because each reflection at the earth involves loss and each passage through the layer also involves attenuation. Thus it will be seen that the most effective rays for long distance transmission are those that leave the transmitting aerial at a shallow angle, and hence an efficient aerial system should concentrate the radiation between about 3° and 15°.

During certain conditions of the ionosphere it appears that a type of transmission is possible which omits intervening ricochets, so that if the initial and final conditions of layer are correct for bending, the intervening condition need only confine the ray to the ionosphere and not necessarily return it to earth.

Reflection from the earth's surface has been discussed on page 94, and it will be seen that at a certain critical angle there may be practically no reflected, vertically-polarised ray.

In order to communicate successfully over a long distance a consideration of Fig. 75 shows that in addition to the attenuation not being too great to permit of the reception of sufficient strength, it is necessary also that the receiving station should not be within the skip zone. This means that the bending in the upper layer must be sufficient to bring down rays within the required distance. The attenuation and the bending of

a ray of a given frequency are dependent upon the state of ionisation of the layers, and this in turn depends on the day-night conditions. We shall expect therefore that to maintain a service it will, in general, be necessary to alter the wavelength used to suit the time of day and season, remembering that the wavelength chosen will generally be that which gives least attenuation with sufficient bending.

Attenuation

This is now not a function of the losses in an equivalent transmission line as it is for long waves, but it is a question of the absorption of a wave as it passes through the ionised layer, and it is found that attenuation is proportional (for daylight conditions anyway) to the square of the wavelength ; in fact exactly the reverse of long-wave attenuation. If the distance of communication depended only on attenuation, the curve for communication distance against wavelength would rise from the critical wavelength as the wavelength decreased. But as the bending effect, which will be considered next, is equally important, we reach a limiting value of shortness of wave for communication on earth, below which the range suddenly falls off to the surface-ray range. This limit is shown on the left-hand side of Fig. 73, where an abrupt drop in the curve is shown at the wavelength of about 10 metres.

This does not mean necessarily that 10 metres will be chosen for communication to the antipodes ; for one thing it is too near the critical value, and secondly, half the earth is never all in intense daylight (when the bending is greatest) and thus one rarely finds in practice any long distance communication carried on at wavelengths below 14 metres.

The amount of attenuation is dependent upon the layer conditions and in general will be proportional to the amount of ionisation. Thus during the day, because ionisation is greatest, attenuation will be greatest, whilst during the night attenuation on all waves is much reduced. For the same reason the attenuation in summer on any given wave will be greater than in winter. Where the route is mainly in daylight, the greatest attenuation will be observed when the ionisation is greatest at a point half way between transmitter and receiver, this occurring about one hour after midday (at that place).

Bending

The bending experienced by a ray of given wavelength depends upon the total change of electron density through which it has passed. The amount of bending with the same change of density varies as the wavelength so that waves below 10 metres are rarely sufficiently bent to return to earth even when the layer is in its most highly-ionised condition.

The ionisation increases quickly after dawn and the amount of bending therefore increases considerably, and on any given wavelength is greatest during the day, gets slowly less during the evening and early night, and is least during the late night period preceding the dawn. Hence the skip zone will be greatest for any given wavelength during the late night and least during the middle of the day. For the same reasons the skip will be greatest during the winter at any given place and be greatest in places of high latitude. Unlike attenuation, however, the condition of layer at receiver or transmitter is more important to bending than the condition between, because the bending into and out of the layer is chiefly concerned with the layer condition at the ends of the routes. Further, that end in the darker grade is really the controlling factor in this respect, for it is no use a ray being bent into the layer if at the other end there is not enough bending to return it, or vice versa.

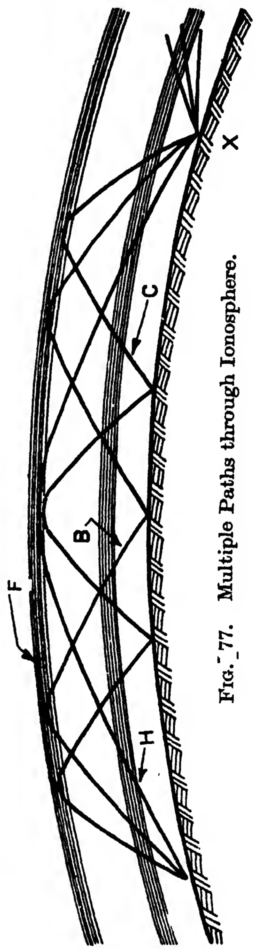


FIG. 77. Multiple Paths through Ionosphere.

The Zenithal Angle

The zenithal angle is of considerable importance, quite apart from the fact that high-angle radiation is lost because it is not returned to earth. The reason for this is that a very narrow pencil of radiation is sufficient to illuminate the whole earth

beyond the "skip" by successive "ricochets," and radiation at other angles only leads to the production of "echo" signals. This is shown in Fig. 77, where *H*, *B* and *C* are three rays leaving the transmitter at different angles such that successive reflections, one of ray *H*, two of ray *B*, and three of ray *C* bring the rays together at a receiver placed at *X*. The different lengths of path of these rays causes fading and blurring of the original signal.

The particular angle which gives the best results at great distances is a matter for consideration and will vary with different circuits, but it is certain that high-angle radiation is of no value and there is strong evidence that the most useful energy is that which leaves the aerial at a very shallow angle, just sufficient to avoid earth losses. Angles approximating to 8° are found best for long distance work, increasing to about 20° for nearer services. Hence the tendency in modern design of short wave aerials is to avoid those forms which give high angle-radiation, such as the harmonic aerial, and to provide systems which propagate energy at a shallow angle, such as the Franklin multiple aerial, which will be mentioned later.

The Ultra-short Wave Band

We have already stated that ultra-short waves are only suitable for short-distance communication and the division between short and ultra-short waves is usually made at 10 metres (30 Mc/s) because reliable long-distance communication is not possible below this wavelength. As has been explained, this is because there is insufficient bending even over the longest all-daylight path possible on the earth's surface. It is evident that, because the degree of ionisation varies so much from place to place and time to time, the shortest wavelength which will be bent back to earth is not clearly defined. Thus although the London Television transmissions on 6·67 and 7·25 metres are only of value for direct-ray communication, occasional strong though usually distorted reception is obtained over very great distances, as, for example, South Africa and America.

An extended series of observations have been made at the Riverhead Station of the R.C.A. Communications Inc.[32] on European ultra-short wave television stations. The afternoon

sound transmission from London was received on most days between September and March and the vision less frequently. The strength of the sound transmission sometimes exceeded 40 db above 1 microvolt per metre. Reception is evidently dependent upon sufficient density in the F_2 layer and the reception conditions appear to follow fairly closely the critical-frequency for the F_2 layer, and when the critical-frequency dropped suddenly in January to low values, signals were not usually received.

Conklin[31] has collected data of amateur 56 Mc/s (5·36) transmissions over distances of 400 to 1,200 miles, it having been found that communication over these distances is frequently possible. These successful receptions appear to depend upon the sudden increases in the electron density of the E layer mentioned previously.

Factors Influencing the Choice of Wavelength

It should now be clear that, broadly speaking, we shall choose for a given communication circuit the shortest wavelength which will ensure rays being sufficiently bent to return to earth within the required distance, since the shorter the wavelength the less the attenuation. As a general rule then, the shortest waves (in the neighbourhood of 15 metres) will be used to cover long distances which are mainly in daylight and longer waves for darkness conditions. The problem is complicated by the fact that many long distance transmission paths are partly in darkness and partly in daylight, and hence the wavelength used must be a compromise. Further, since conditions are continually changing, more than one wave may be necessary to maintain any particular service throughout the 24 hours.

From what has been said, it is clear that the prediction of the performance of a short wave circuit, or the choice of a suitable wavelength for a given service, is a very difficult matter. A variety of charts have been produced from time to time, to enable the information to be obtained for average conditions, and these charts generally take one of three forms.

(1) The ionosphere conditions for a given number of seasonal conditions and for all latitudes may be charted, so that with the correct season-chart and a complementary map of the world, it is possible to observe the conditions over any route

and at any time. The chart showing the ionosphere conditions will usually be made transparent, so that by sliding this over the map to the correct position for the time of day, the general ionosphere condition on the great circle route between the two places can be observed. Then by reference to " skip " curves and charts for field strengths for the different layer conditions, it is possible to estimate the probable field strength for a given power and distance. Such a system is, of course, universal in its application, but experience is needed for correctly interpreting the information so derived for those routes which pass through several grades of electrons density in the ionosphere.

(2) When one is concerned only with transmission from (or reception to) a given place, it is possible to produce a series of charts to show the best wavelength to use at any time and season, for communication with a place at any latitude and at any given distance. Such a set of curves is simple to use, but as they refer to only one place, they are not universal in application, and to make them so would necessitate the production of a very large number of sets for the different latitudes.

(3) It is possible to produce a comparatively simple series of curves for giving the field strength between stations of given latitudes and for a limited range of distances, not too great or small.

We propose to indicate the basis on which such sets of curves are produced, and in Appendix II will be found a series of curves suitable for general work.

(1) Eckersley and Tremellen Charts

Although recent charts are produced in a different form and will be described later, the charts as originally produced were in a pictorial form which lend themselves better to a general explanation of the principles on which charts of this type are based.

Three charts, to cover equinox, summer and winter seasons, are shown in Figs. 78, 79 and 80. In these charts it will be observed that the infinite shades of ionosphere conditions have been limited to six grades, A, B, C, C–C^1, C^1 and D, these grades corresponding to : (A), conditions of intense sunlight ; (B), weak sunlight ; (C–C^1, C^1), twilight ; and (D), darkness. These various grades are shown labelled in the Figs. 78, 79 and

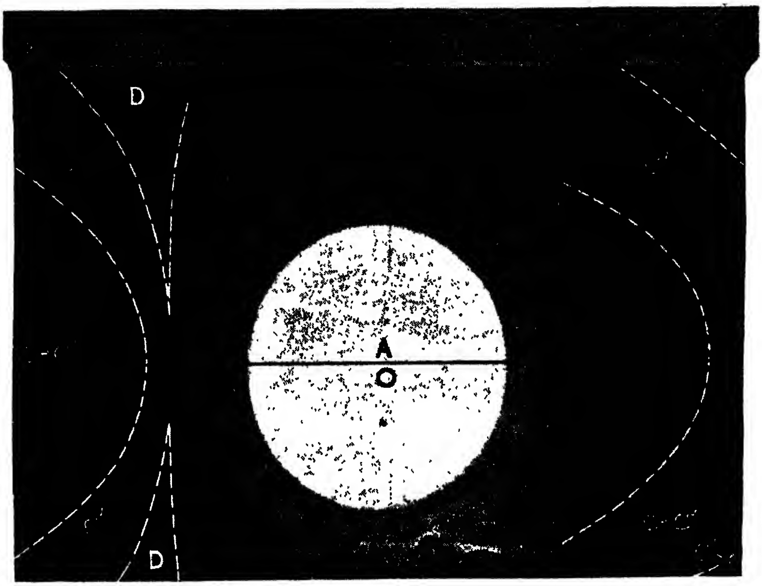


FIG. 78. Equinox Chart.

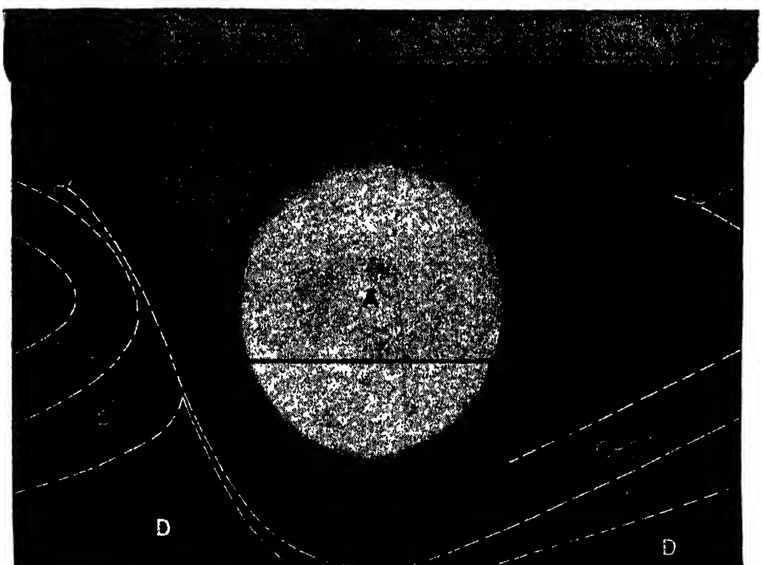


FIG. 79. Summer Chart.

80 and these grades A, B, C, C–C^1, C^1 and D correspond to the six grades 1 to 6 in the skip curves of Fig. 76. The sunset-sunrise line is also indicated by a thick black line, which takes the form of two vertical lines at 6 a.m. and 6 p.m. at the equinox period and of " cosine-shape " curves during winter and summer.

The fact that—although the E and F_1 layers follow approximately the sum's altitude—the F_2 layer behaves in a contrary

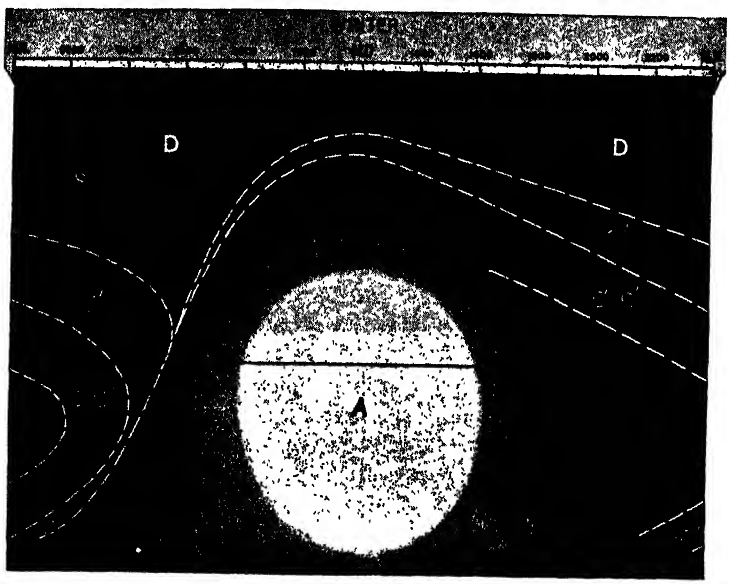


FIG. 80. Winter Chart.

fashion, presents difficulties in the production of such charts, but the average behaviour suggests the necessity for the lagging effects of the darker grades in the charts as shown. It may help to explain the utility of such charts if the transmission characteristics of the various grades are enumerated.

Transmission Across the All Daylight Zone. The longest distance on the earth which can be said to be all in intense daylight at any one time is about 6,000 miles, and as the attenuation over such a daylight path will be high, the shortest possible wavelength will be used. This will be about 14 or

15 metres, as shorter waves than these would have a skip distance exceeding 6,000 miles. A shorter distance would necessitate an increase in wavelength with a corresponding increase in attenuation, and the great attenuation of this wave will prevent the communication extending much beyond this distance. The minimum signal will occur when it is a little past mid-day at a point halfway along the path.

Transmission Across the Twilight Zone. Across the twilight zone the attenuation on all wavelengths is much less than for a daylight zone and for wavelengths below 20 metres becomes especially small. When the seasonal condition and time of day is such that the great circle between the stations is near the dividing line between daylight and darkness (called the shadow line), the attenuation is so small that signals may make more than one journey round the earth, and produce what are termed "round the world echoes" at the receiver. For at this time the ionosphere condition along the great-circle path between the stations is the same around the whole earth, and hence it is easy to find a wavelength exactly suitable, whereas if the great circle line lies across different grades of layer condition, we cannot find a wave which will pass through these varying areas, for any wavelength which is long enough to pass through the darker grades will be rapidly attenuated in passing through the daylight zone. It might be thought that it is not possible for two stations to be in the all twilight zone except for a brief period each day, but a reference to the charts will show that what is there termed "twilight" includes weak sunlight such as experienced in high latitudes in winter. For instance, during an English winter we are never in "daylight" within the meaning attached to it in the charts, and when working to other stations in the northern hemisphere the great circle line will be either in a twilight grade or a darker grade throughout the 24 hours.

We may say that waves suitable for routes in the twilight grade extend from 14 metres up to about 40 metres, the longer distance services utilising the shorter wavelengths.

Transmission Across the Darkness Path. Across this zone between the twilight and the late dark area the attenuation on wavelengths of the order of 20 to 60 metres is very slight, but on waves about 20 metres the signal strength would

commence to fall off rapidly, not because of the attenuation, but because of insufficient ray bending.

Transmission Across the Late Darkness Zone. For the portion of the earth which may be considered as in late darkness, waves of greater than 50 metres are suitable, and 30 metres represents a critical wavelength below which bending will usually be insufficient. Waves less than 30 metres may pass successfully through a short length of late darkness path, if it is intermediate between the stations and not over one or the other. This to some extent can be understood, because, as previously mentioned, if the conditions at the commencement and the end of the ray path are sufficient to create a bending of the ray into and out of a path coincident with the layer, then the intermediate condition is concerned chiefly with attenuation and not bending.

The type of charts now being used are not divided into arbitary grades of darkness as just described, but in contour lines of critical wavelengths for oblique reflection. This is, of course, essentially the same, since (as we have already seen) there is a direct connection between electron density and critical-wavelength. The dividing of the ionosphere into such critical-wavelength contours, however, gives a direct indication of the minimum wavelength to which the layer is not transparent over any given great-circle path that is being considered.

(2) Field-strength Contour Charts

An alternative method of producing charts has been suggested which aims to map contour-lines of field strength from any point of given latitude, for different wavelengths, seasons, and times of the day. Such a system would be an ideal one, but unfortunately a prohibitive number of charts is necessary, since, not only must a chart be provided for each hour of the day, and the various seasons, but also for every 10° of latitude, and, of course, for a large number of wavelengths. This would necessitate about 1,400 charts to cover the short wave-band.

Such a chart is shown in Fig. 81, and this is given as it is illustrative of the conditioning features of short wave transmission. The chart shows a series of contour lines of equal field-strengths from a 1 kW transmitter, situated in England

(latitude 52°), the time being midday, in winter, the contours being denoted in db, plus, and minus from a zero constant of 1 μ V per metre. The small shaded area in the centre is the

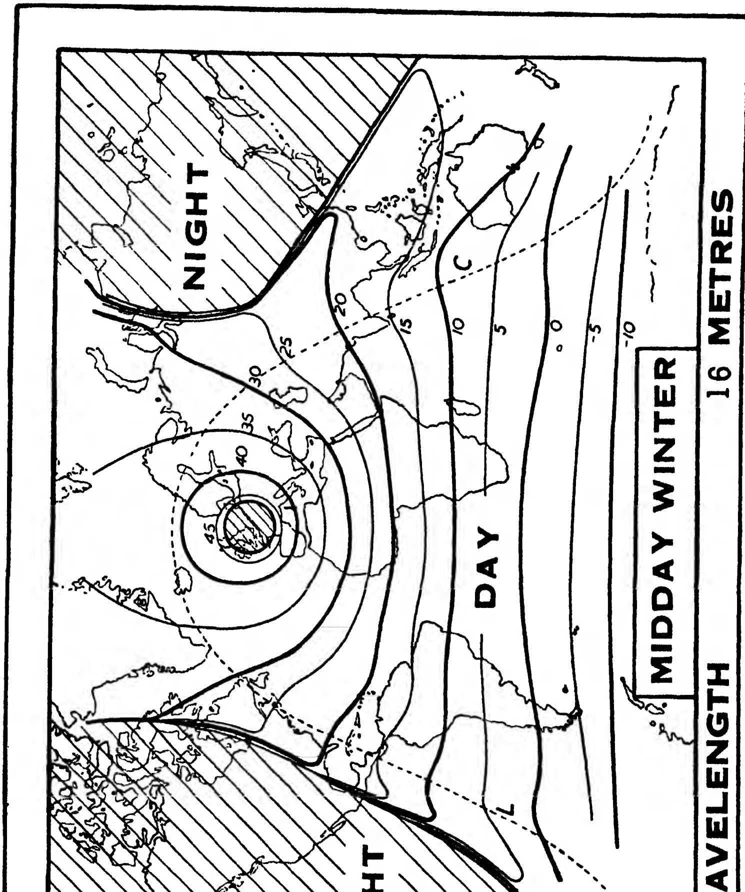


FIG. 81. F-S Contour Map.

skip area, after which the ionosphere rays are returned to earth. From the skip area the rays follow a ricochet course until they encounter the dark grade (shaded), which has not

sufficient electron density to return them to earth again. Thus from the skip area to this dark area we have contour lines of equal field strength, spreading outwards. Towards the south the rays pass through an all daylight grade where the attenuation is highest and is the conditioning feature. To the east and west the daylight is less intense and attenuation in consequence less, and in regions near the shadow-band (shown dotted) is very small indeed. This causes the contour lines to be drawn out towards the shadow line, so that high field strengths are obtained at very great distances from the transmitter.

To the N.E. and N.W. the direction is into the darkness area and the possible communication distance is thus shortened and, in fact, signals cease almost abruptly, owing to the lack of bending, because the region of critical density has been reached.

Charts for Field-strength, Distance, for Different Latitudes

Yet another type of chart, which is useful for a limited range of distances only, has been developed, which gives the field-strength direct, for different seasons, and times of the day, and between stations of different latitudes. A series of such charts is given in Appendix II.

It now remains to discuss some peculiarities of short wave propagation, the principal being the fading phenomena.

There are two distinct types of fading giving the following effects:

(1) Complete cessation of communication for many hours.

(2) Variations of signal strength, the change of which may be slow or very rapid in character.

Complete Fading

Complete fading is liable to occur if the transmission path lies in high latitudes, and particularly if it passes near the magnetic poles, when it may be so bad as to eliminate signals for several hours. In one or two cases signals have disappeared for as long as one or two days, but usually these periods will be of short duration, and there will often be only a general reduction of the signal level and not a complete cessation. These complete "fade-outs" occur during magnetic-storm activity, and such are often associated with sun-spots, which have a cyclic variation of about 27 days, the time of the sun's rotation.

Systematic observations of the correlation of fading with sun-spots and magnetic storms have been made by T. L. Eckersley over a number of years, the sun-spot activity being observed by the simple expedient of pointing a focussed telescope to the sun, holding a chart a few inches below the eyepiece at such a distance that the sun's image is two or three inches in diameter, and plotting on the chart the position of the sun-spots or sun-spot patches. These vary considerably in size and number, but all those of interest are quite large enough to be observed by this means. It is found that those spots which pass exactly through the centre-meridian of the sun may have the most effect, and this is to be expected if the spot is a hole in the sun's envelope. Secondly, the influence on wireless signals appears from one to three days after the sun-spot has passed the centre-meridian, so that whatever is causing the fading is some agent which travels at a much slower speed than light. A further point is that the size of the sun-spot is no criterion of the effect produced, and many periods of great sun-spot activity may pass without any marked effects on wireless circuits. Sun-spots and magnetic storms, besides having a period of activity recurring each month, have also a long cyclic variation, the periods of peak activity recurring every 11 years, 1939 being the last. Again, stations in high latitudes are chiefly affected by magnetic storms and sun-spots, and therefore it is assumed the agent causing the disturbance is divided by the earth's magnetic field, towards the magnetic poles.

The fading has been attributed to increased attenuation due to the greater ionisation. Extensive tests on ultra-short waves were tried over a long-distance circuit during these abnormal times on the assumption that if the ionisation is so much greater the bending might be sufficient to bring them down to the earth's surface within the required distance. Except for occasional loud signals, no useful results were obtained, and there is no method yet known of overcoming this type of fading.

It should be mentioned that in addition to causing fading, the magnetic-storm activity raises the level of ionisation, particularly at high latitudes. This means that the correct wave-length to choose for any channel in high latitudes will vary

cyclicly with the magnetic storm era, and as an illustration we might mention that the circuit between this country and Canada was forced to employ waves as long as 60 to 70 metres at night during 1933, whereas it used 30 metres during the maximum magnetic-storm period of 1939.

Catastrophic Disturbances

In addition to complete fade-outs being caused by magnetic-storm activity, a rare disturbance has been observed which affects all daylight routes. This disturbance, which may last for periods of a few minutes up to half an hour, is not definitely associated with magnetic storms, but has somewhat similar effects (except that they are world wide). It has been observed that these disturbances are associated with hydrogen eruptions in the sun's chromosphere.

Rapid Fading

It is usually considered that the causes of this type of fading are ray interference and change of wave polarisation.

Consider ray interference. As was seen in a previous section, a long-distance, short wave signal is not a single ray, but may be made up of two or more rays arriving by different paths. If we have only two rays, the phase of these will determine the resultant field at that point, which can vary from zero, if the rays are equal and completely out of phase, to their sum, if in phase. If more than two rays are involved the resultant depends upon their vector sum.

It is found that even if the receiver is in such a place that it cannot be energised by ricochet interference, fading can still occur. In this case the interference is between closely adjacent rays following approximately the same path. If the ionosphere were a stable medium merely graded vertically in a definite unchanging manner, then these contiguous rays would " illuminate " uniformly a small area of the earth's surface. As, however, the ionosphere is not homogeneous either horizontally or vertically, the different rays forming the ray-pencil are disarranged in their passage through it and now " illuminate " the area in a non-uniform fashion. As the ionosphere conditions vary from instant to instant, the distribution over the

area considered is continually changing and hence fading of the signal results.

Actually the variation of signal strength is greater than a calculation giving the resultant of a number of vectors having random phase relationships would lead us to expect. This is to be accounted for by an additional variable factor, namely, changing polarisation of the waves. Thus if we use a vertical aerial for reception it means that only vertically-polarised waves, or those waves giving some vertical component, will be received. Hence a single wave, the polarisation of which changes with time, will produce a variable signal apart from any interference variation.

It is of interest to note that fading is different in time at points quite near together in space. Thus we can, to some extent, overcome fading in the following ways:

(*a*) By using both vertical and horizontal aerial systems.

(*b*) Summing the energy received on a number of aerials spaced apart a sufficient distance.

Very deep fading is frequently found when a station is near the edge of the skip distance. This is due to the fact that very small changes occurring in the ionosphere make considerable changes in the direction and intensity of the rays coming down near the station, at one moment the rays falling between the transmitting and receiving stations, now passing over the latter.

At any one place the fading is different for different frequencies separated by only a few cycles. If fading is caused by interference between rays following different paths this will readily be understood, for the bending of a ray is a function of its wavelength and quite a small difference of wavelength (and therefore bending) will change the points at which a group of rays again meet. Thus a continuous-wave signal suffers more violent fading than a modulated wave because the latter involves the transmission of a wide spectrum (see Chapter III), and there is a better chance of collecting part of the signal in this case, because the individual frequencies traverse different paths. For this reason telegraph transmitters are frequently modulated, so that the energy is conveyed to the receiver by a band of frequencies instead of a continuous wave being used. Although a modulated carrier may suffer less fading, such a signal is often received distorted because the component

waves are altered in relative value. This is sometimes referred to as selective fading. This does not matter for telegraphy, but may be serious for telephony or facsimile telegraphy. If the fading is due to a number of rays adding together at random phase, the distortion is found to be serious and when the distortion is small it is probably because the fading is chiefly due to changing polarisation. Yet another type is characterised by fading of the modulated component but no apparent fading of carrier wave. One of the authors has suggested * that this type of fading is due to a phase-shift of carrier relative to the side bands as the wave passes through the ionised layer, for it has been shown in Chap. 3 that a phase shift of carrier of 90° relative to its correct phase can almost eliminate amplitude modulation although a frequency modulation is substituted. In short wave communication, fading is generally the most difficult feature which has to be overcome, as it is to be found at all times, on all wavelengths, and it is this feature which necessitates an average level of signal strength for commercial circuits much above the noise level of the circuit.

Modern technique is gradually overcoming the ordinary fading phenomena and an up-to-date receiver can deliver a constant level of signal at nearly all times, of a quality good enough for commercial telephony.

Echoes

It is found that when a single short wave signal is transmitted, more than one signal is sometimes picked up on a distant receiver. It has become customary to speak of all the received signals subsequent to the first as "echoes," though these signals are not produced in the same way as an ordinary echo, i.e. by a wave reflected from a large object and returning back more or less along the path of the incident wave.

Although echoes have actually been heard on the longer waves, it was the short waves which first called attention to their existence in wireless telegraphy, and in general they may be classed under three headings :

(1) Very long delay echoes.
(2) 1/7 second echoes.
(3) Quick echoes.

* Marconi Review, No. 23.

(1) Very Long-delay Echoes

Long-delay echoes, called the Stormer echoes, after Carl Stormer who first observed them, may appear as long after the signal as 10 seconds (30 seconds has been mentioned) very much distorted. In fact it is because the echo is so unlike the original signal and so long after it that its existence was unobserved for a very long time after short waves had been a common means of communication. Naturally an echo with such a delay is very difficult to account for, as at the velocity of light the echo signal would have to travel hundreds of times round the earth, and theories put forward are various in number. There is no scope within the present book to set out these theories, particularly as the long-delay echo is not of great interest to the wireless engineer, and none of the theories put forward so far have been established. It should be pointed out that the Stormer echo is such an extremely rare phenomenon that it has only been reported by a few observers in Europe and Asia.

(2) 1/7 Second Echoes

These echoes, the " all-round-the-world " echoes, are caused by a signal making a complete circuit of the earth. Thus a station receiving from the west, say, may get an echo signal from the same direction approximately 1/7 second later than the original signal, should the condition of the layer be favourable. Although weaker than the original signal, it may be at times troublesome, as a directional system cannot eliminate it. The times at which these echoes are heard usually coincide with the period of season and day when the great circle line between the stations is nearly coincident with the shadow band, and wavelengths between 15 and 18 metres show this type of echo in a very pronounced manner.

" All-round-the-world " echo is a common phenomenon on a local station at the right season and time, but a more rare phenomenon on long distance working. This will be understood by studying the shadow charts ; for a local transmitting station is in the same grade as the receiving station and at periods when the stations are near the shadow band the signals will follow the shadow band great circle path. On the other hand if a long-distance station is considered, there is only a short interval of time at one particular season which can

possibly create echo; further, the attenuation of the signal along the extra distance to be covered between transmitter and receiver makes the chances of echo less.

In addition to the "all-round-the-world" echo, a shorter-interval echo will often be heard at the same time, which is caused by a portion of the transmitted energy taking an opposite path round the earth and so arriving at the receiving station from a direction approximately opposite to that of the main signal and at a time interval dependent upon the positions of transmitter and receiver. This echo is not troublesome if directional systems are used as they will usually eliminate it or render it too weak to have any serious effect. These two echoes are always clearly defined and there is no appreciable distortion.

(3) Quick Echoes (Multiple)

Quick echo is made evident in different ways; thus when a tone or speech is being received, a blurring or distortion of the signal may result or it may have a "hollow" sound giving the ordinary "empty room" effect. Signal blurring and distortion are due to multiple ray reception and as the distance increases their effect becomes less marked. This is explained by the fact that the rays making the longest "ricochets" are less attenuated, for at each ricochet from earth and at each passage through the E layer a little energy will be lost and thus at very great distances there may be only one really strong ray left, in consequence of which the echo signal will be negligible. This quick or multiple echo is so close behind the main signal that when recording telegraph signals it shows up merely as a lengthening of the "marking" periods as if the relays had been given a "marking bias." For this reason the effect was known as "ether bias" in the early days of short wave working, when the reason for it was not clearly understood. Except that it may cause fading, this multiple echo does not interfere with telegraph reception, as it may be corrected by the use of suitable "shaping circuits." The introduction of facsimile showed the effect up in its true light, namely as a series of separate signals, an example of multiple in facsimile being shown in Fig. 82, and measurements since made show that the blurred signal is in reality a number of signals arriving by paths of different

lengths, one milli-second being the greatest time interval yet observed between individual rays. By careful check of these multiple echo times it is but a geometrical matter to arrive at the number and length of each ray path, from which an estimate can be made of the height of the ionosphere.

In addition to multiple rays giving the quick echo, scattering has also some part in the various multiple effects observed,

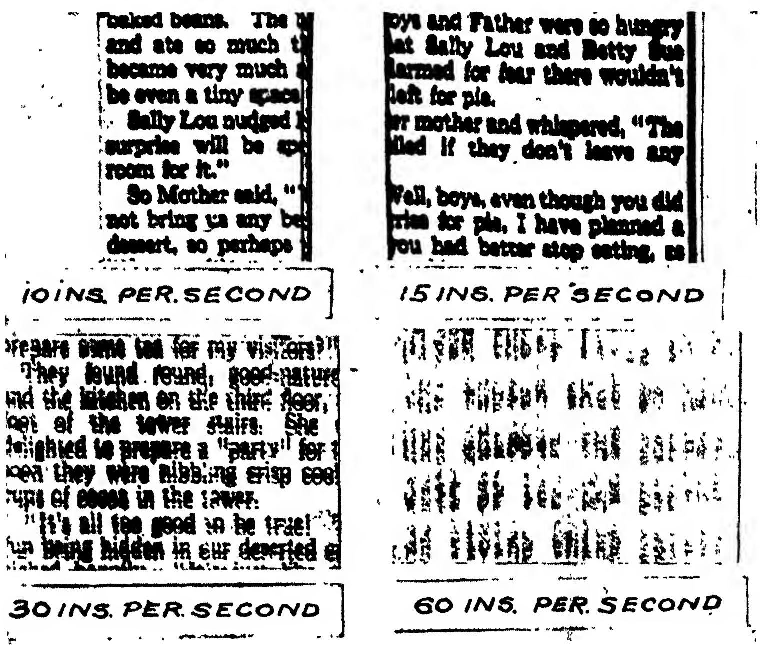


FIG. 82. Effect of Multiple Fading on Facsimile Reception.

for if the scattered radiation is present at a strength comparable to the main signal this slightly modifies the sound heard and the blurred signal is most probably due to both multiple and scattered signals combined. The " empty-room " effect which is unmistakable when heard, is generally considered to be due to scattering, the scattered signal providing a background to multiple signals which gives an effect very closely analogous to the sound background in an undraped empty room.

Atmospherics

It is customary to imagine that atmospherics are negligible on short waves. Actually they are quite strong on the longer short waves during the summer months, and it is only on waves below 15 metres that they can be said to be quite negligible. But on waves of this order we have considerable increase of general noise level due to machines, magnetos and loose mechanical contacts.

Atmospherics are, of course, electric disturbances in the lower air layers caused by lightning discharges. The effect of atmospherics on communication is not so much dependent upon the intensity of any single atmospheric, as on the frequency with which they occur. Because the power involved is so enormous, an atmospheric centre thousands of miles away can create serious interference. Thus wireless services conducted by stations in temperate climates can be more upset by atmospherics emanating from zones near the tropics than from local storms, because the latter are infrequent whereas the former are almost continuous.

Most of the atmospheric-producing centres are large land areas near the tropics such as Africa, Northern Australia, Northern South America and India. These atmospheric centres do not remain stationary but vary periodically with the sun, moving some 10° north during our summer and some 10° south during our winter. In addition to the seasonal change of position of the atmospheric centres, the actual amplitude and frequency, particularly frequency, vary with the sun's altitude diurnally, their frequency reaching a maximum at 3 p.m. local time.

The atmospheric pulses produced are very varied in character, but since their wavefronts are usually steep they are capable of giving interference on an infinite spectrum of waves (since a pulse is an infinite series of periodic frequencies), and thus the atmospheric is to be treated as an interfering transmitter radiating power at all frequencies.

At 100 metres, atmospheric interference is extremely bad in tropical zones, and down to 30 metres the interference can be serious at times, at all places. It is to be observed that stations in a temperate zone may suffer more from atmospherics on the very short waves than stations in the

tropics very close to the atmospheric centre, for the tropical station will be inside the "skip" area and thereby unaffected.

Precipitation Static

It has long been known that when an aircraft flies through rain, snow or dust, a loud rushing noise may be heard in the output of receivers in the aircraft and this may be of sufficient volume to prevent the reception of any signals. Unfortunately, it is in just such times of storm that it may be most important for the safety of the aircraft that contact should be maintained with ground stations.

This interference, known as "precipitation static," has recently been the subject of a large-scale investigation in the U.S.A.[35] It has been shown that an aircraft flying through particles becomes charged by friction, the charging current sometimes being sufficient to increase the potential of the aircraft at a rate of 200,000 volts per second.

As the potential rises, however, corona occurs from any points on the aircraft, thus discharging it. If any of these discharging points are near the aerial, interference with reception is caused. The most severe (and quite usual) condition is when the aerial itself is one of the discharging points.

If an aerial is covered with sufficient insulation, no discharges can take place from it. The mechanical and electrical requirements of a suitable insulator are severe but have been met sufficiently well by polyethylene.

In addition, aircraft may be fitted, at positions as remote from the aerial as possible, with devices for discharging the aircraft (such as points or wicks) with as little "fuss" as possible.

Measurement of Received Signal Strength

In the foregoing sections an outline has been presented, making use of various theories and assumptions regarding the ionosphere, etc., and certain methods of investigating ionosphere properties have been given. We will now discuss one or two other measurements associated with the subject.

One of the most important and practically useful measurements in connection with any investigation into the propagation of wireless waves must be that of received signal strength.

Even on long waves this presents considerable instrumental difficulties if weak signals are to be measured, because the incoming power is insufficient to measure by direct means.

When measuring short wave signals grave difficulties are introduced by the nature of the signal, since this is rapidly varying in strength and is usually polarised in a complex fashion. In addition, the high frequencies involved make the design of accurate apparatus very difficult, as extremely small accidental capacity couplings may so greatly modify results.

Practically all signal strength measuring apparatus depends upon matching the signal against a known output from a local source, using the same receiver for both signal and local source, one of the most important points in design being the careful screening of the latter. It is usual to employ, for reception, a frame aerial, as by this means the magnetic field in all directions can be measured and also the effective area (which must be known in order to express results in microvolts per metre) can be determined better than the effective height of an open aerial.

Measurement of Received Wave Polarisation

Appleton and Ratcliffe examined the downcoming wave of a south to north transmission at broadcast frequencies and found it to be circularly polarised with a counter-clockwise sense of rotation, whilst a corresponding experiment conducted in Australia showed a clockwise rotation, thus demonstrating that the circular polarisation is produced by the earth's magnetic field. Eckersley has obtained simultaneously oscillograms of the same signal received on a horizontal and a vertical aerial, and found that in some cases the fading on the two is opposite in phase, thus indicating an elliptically polarised wave.

The cathode-ray direction finder of the Radio Research Board has been employed to examine the polarisation of short wave reception. The signals from two loops at right angles to each other are put through identical receivers (adjusted to give the same performance) and applied to the two sets of deflecting plates in a cathode-ray oscillograph. The arrangement is therefore similar in theory to a Bellini-Tosi direction-finder with the cathode-ray oscillograph as the goniometer,

and a vertically polarised surface wave such as is produced in long wave transmission would give a straight line at an angle which would indicate the bearing of the transmitter.

When the arrangement is applied to short wave signals, however, an ellipse is usually produced, due to the presence of horizontally-polarised components, though a rotating straight line is sometimes observed. The ellipse is continually varying both in direction of axes, in shape and in size.

It is evident that further extended study of the nature of the received signal at long distances is of considerable practical importance owing to its probable influence on the design of receiving aerial systems.

CHAPTER VI

RADIO-FREQUENCY LINES, OR FEEDERS

IT is the usual practice nowadays to install several transmitters (or receivers) in one building and erect the aerials or aerial arrays for them some distance away. Hence radio-frequency transmission lines or feeders become necessary between the aerial and transmitter (or receiver), the function of which is to convey high-frequency power with the minimum of loss. In the case of an aerial array, the feeder system will also have to be arranged to supply currents having the correct phase relationships to the individual aerials. The lines take the form either of parallel wires or concentric tubular conductors.

Adjustment of Load Impedance for Maximum Power

It may be useful here to deal with an important general condition to be realised in all telecommunication circuits if we are to get the greatest possible output from them, namely, the principal of matched impedance. Suppose that we have a resistive load R_o, supplied from an alternator, having an internal resistance R_g and internal reactance X_g (Fig. 83).

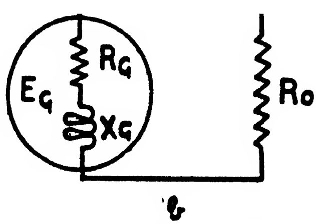


FIG. 83. Impedance Matching.

The total impedance is $\sqrt{(R_g + R_o)^2 + X_g^2}$

and hence current I is equal to $$\frac{E_g}{\sqrt{(R_g + R_o)^2 + X_g}}$$

and output W is $$\frac{E_g^2 R_o}{(R_g + R_o)^2 + X_g^2} \quad . \quad (1)$$

The value of R_o which makes W a maximum will be that which makes $\frac{dW}{dR_o}$ equal zero.

Differentiating W by the quotient rule we have

$$\frac{dW}{dR_o} = \frac{(R_g^2+2R_gR_o+R_o^2+X_g^2)E_g^2 - E_g^2R_o(2R_g+2R_o)}{(R_g^2+2R_gR_o+R_o^2+X_g^2)^2} \quad (2)$$

Hence W will be a maximum for that value of R_o which makes

$$R_g^2+2R_gR_o+R_o^2+X_g^2-2R_oR_g-2R_o^2=0$$

from which $$R_o = \sqrt{R_g^2+X_g^2} \quad . \quad . \quad . \quad . \quad (3)$$

It will be seen that the maximum power is transformed to the load when its resistance is equal to the internal impedance of the alternator or other source and the load and source are then said to be matched.

If the load has also reactance but the phase angle of the load is constant (in other words, the reactance bears a constant ratio to the resistance) then a similar analysis to that given above will show that for maximum power transferred to load, the magnitude of the load impedance should equal that of the generator impedance. Since an ordinary transformer has the effect of transforming the magnitude of the load impedance, whilst leaving the phase angle practically unchanged, this is an important practical case, especially in low-frequency work.

In the general case where the reactance of the load can be varied independently of the resistance, it can be shown that when the load resistance equals the generator resistance and the load reactance is equal in value to the generator reactance, but of opposite sign, then the power transferred to the load is the maximum possible. Since the reactances are equal but of opposite sign, it is evident that the circuit is in resonance.

It should be understood that the matched impedance condition does not necessarily give high efficiency. For example, if generator and load have only resistance, then the efficiency is 50%.

This condition of matched impedance is continually being met with throughout telecommunications, whether generator or load are in close proximity, or whether they are separated by a feeder line. The only difference is that in the former case matching of load to generator automatically ensures

maximum output, whereas in the latter case we have to match load to line to minimise line losses, and load and line to generator to obtain maximum output from the latter.

In many cases the source and load impedances will be fixed from other considerations, and it will then be necessary to insert some transforming device. In the case of voice-frequency work an ordinary transformer would be used with a suitable ratio of primary to secondary turns, but a transformer of conventional type would be very difficult to design for high frequencies, due to the magnitude of stray magnetic leakage and capacity effects, and, moreover, as will be seen later, simpler solutions are available.

Electro-magnetic Waves along Lines

Fig. 84 gives a picture of what we may suppose is occurring when a line is connected to a generator. The dielectric between the two conductors will be traversed by electric

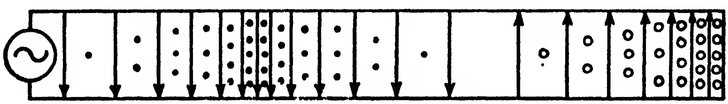


FIG. 84. E-M Wave on Lines.

strain lines stretching between the two conductors. Since these lines are in motion, magnetic lines will be produced, and these will be concentric with the conductors. In the "skin" of the conductors there will be a drift of electrons, or, in conventional terms, a current. No energy is required to maintain an electro-magnetic wave in a pure dielectric, but the current in the guiding conductors results in some energy being converted into heat (as in all electric conduction) and, in addition, there must be some loss in the solid insulation necessary to support the conductors. This loss results in a reduction of the velocity with which the wave travels down the line, whereas if there was no loss the "guided" wave would travel at the same speed as a "free" wave, i.e. at very nearly 3×10^8 metres per sec when the dielectric is air and $\dfrac{3 \times 10^8}{\sqrt{\kappa}}$ when the dielectric constant is $\kappa\kappa_0$. In the case of lines carrying

very high-frequency currents, the energy loss is fairly small compared with other effects and the reduction of velocity is small. Hence as a fair approximation a working theory can be developed neglecting loss.

Correct Termination to Prevent Reflection

Assume now that the wave reaches the end of a line which is open-circuited (or short-circuited), then since there is no circuit to accept energy there will evidently be reflection, and a wave will travel back along the feeder towards the generator. Here there may be a further reflection, so that the actual distribution of electric and magnetic lines is the resultant of a number of waves travelling in both directions along the line.

Thus a resultant forward wave will be produced which is the vector sum of a number of successive forward wave components, each of smaller amplitude than its predecessor, because of line losses; and a corresponding reflected wave will also be formed of the vector sum of a number of reflected wave components of diminishing amplitude, the vectors generally taking the form of a spiral, in which case the resultant waves, forward and reflected, will not be unduly large. But should the length of line be such that each reflection is in phase with other waves travelling in the same direction, a very large resultant forward wave (and reflected wave) will be formed many times greater in amplitude than the original wave impressed from the generator. In fact in these circumstances it is only the line loss which prevents these resultant waves becoming infinitely great, as can be imagined by considering the sum of an infinite number of vectors of equal lengths and in time phase.

If now two plane waves, both of the same amplitude, each travelling in opposite directions along the same line, are combined, they will be found to form a stationary wave. In all cases such a stationary wave will consist of current and voltage waves, the nodes of which will be spaced a quarter-wavelength apart, the exact positions of these waves on the line being determined by the terminating conditions. In the case of a line terminated by a loss-free circuit, there is no transfer of energy along the line at all (if its conductor resistance is neglected), but only a surging to and fro of energy. The

reader will be familiar with stationary waves as exhibited on strings, etc. If a line is terminated by a dissipative circuit, there will be a forward travelling wave supplying the power from the generator. In addition, there may be repeated reflections of energy forming a stationary wave.

The presence of the stationary wave is undesirable because it reduces the amount of energy actually transferred by the feeder for a given generator voltage and also increases the losses because the current in the conductors is increased. We need to know, therefore, how to prevent reflection.

At the far end of the line the ratio of voltage to current is fixed by the impedance of the circuit placed there. If then the electric and magnetic components of the advancing wave do not satisfy this relationship, a reflected wave must be formed such that the vector sum of advancing and reflected waves add to give the required condition at the load. The remedy for reflection is, therefore, to provide a circuit whose impedance requires just the relationship between electric and magnetic fields which actually exists in the original advancing wave, and, as will be shown later, this relationship depends upon the line constants. Thus the terminal load should be such that it absorbs the energy at precisely the rate at which it arrives.

Another way of considering the matter is to say that we require an output circuit which, while it may be very different in appearance from the feeder, is from the point of view of the advancing wave the same, so that there is no apparant change of medium and therefore no reflection.

Equations for Current and Voltage along a Line with Negligible Losses

When we wish to analyse the behaviour of lines mathematically it is much more convenient to regard the line as made

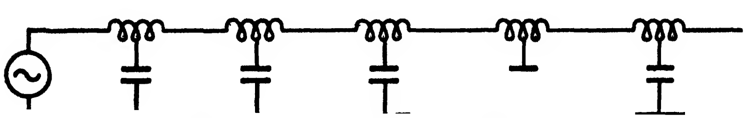


FIG. 85. Equivalent Circuit of Loss-free Line.

up of distributed inductance and capacitance, instead of speaking of electric and magnetic fields in the dielectric. This is only a difference of nomenclature, of course, and not a phy-

sical difference, since when we speak of inductance we mean flux linkages per ampere and when we speak of capacitance we are expressing electric strain lines per volt. Our line can, therefore, be represented by the circuit of Fig. 85, the inductance per unit length being denoted by L and the capacitance by C. Actually the line is more nearly equivalent to the circuit of Fig. 86, because the conductors possess resistance and there

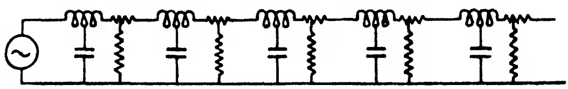


FIG. 86. Equivalent Circuit of Line with Loss.

must be leakage across insulators and dielectric loss in them. If we applied voice frequencies to our line, then we should have to take these into account even in an approximate treatment, but at high frequencies the inductance and capacitance effects are so great compared with the resistance and leakage that the two latter are much less important. Thus a simplified analysis,

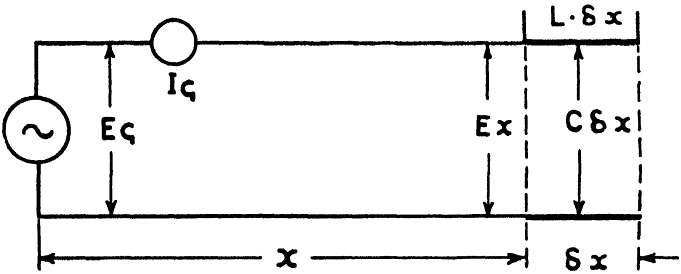


FIG. 87. Illustrating Line Equations.

assuming no losses in the line, will be developed to show the general working of radio-frequency lines.

It should be clearly understood that the actual losses, when a line is used at R.F., will be much greater than if it is used at telephone frequencies, but the resistances have much less effect on the voltage and current distribution; also in R.F. we are usually dealing with comparatively short lengths of line.

Consider an infinitely long feeder (Fig. 87) supplied from an alternator giving a sinusoidal E.M.F. E_g. Then consider any point at a distance x from the alternator and let the maximum

value of the voltage at x be E and maximum value of current at x be I.

Then *decrease* of voltage over δx (assuming current constant over this length)

$$= j\omega LI\delta x$$

or
$$\frac{dE}{dx} = -j\omega LI \quad . \quad . \quad . \quad . \quad (4)$$

Decrease of current over δx (assuming voltage constant over this length)

$$= j\omega CE\delta x$$

or
$$\frac{dI}{dx} = -j\omega CE \quad . \quad . \quad . \quad . \quad (5)$$

Differentiating (4) we have

$$\begin{aligned} \frac{d^2E}{dx^2} &= -j\omega L\frac{dI}{dx} \\ &= (-j\omega L)(-j\omega CE) \\ &= -\omega^2 LCE \quad . \quad (6) \end{aligned}$$

Similarly
$$\begin{aligned} \frac{d^2I}{dx^2} &= (-j\omega C)(-j\omega LI) \\ &= -\omega^2 LCI \quad . \quad . \quad . \quad . \quad . \quad (7) \end{aligned}$$

These differential equations are of a well-known form, having solutions

$$E = A\cos mx + B\sin mx \quad . \quad . \quad (8)$$

$$\text{and } I = C\cos mx + D\sin mx \quad . \quad . \quad (9)$$

where $m = \omega\sqrt{LC}$ and A, B, C and D are constants.

When $x = o$, $E = E_g$

$$\therefore E_g = (A \times 1) + (B \times o)$$

or $A = E_g$.

Differentiating (8)

$$\frac{dE}{dx} = -mE_g\sin mx + Bm\cos mx$$

when $x = o$, $I = I_g$ and $\therefore C = I_g$.

Differentiating (9)

$$\frac{dI}{dx} = -I_g m\sin mx + Dm\cos mx.$$

But $$\frac{dE}{dx} = -j\omega LI$$

and hence

$$-mE_g \sin mx + Bm \cos mx = -j\omega L\,(I_g \cos mx + D \sin mx).$$

As this is true for any value of x it follows that sine and cosine terms must be equal.

$$-mE_g = -j\omega LD$$

$$D = \frac{mE_g}{j\omega L} = \frac{\omega\sqrt{LC}\,E_g}{j\omega L} = -jE_g\sqrt{\frac{C}{L}}$$

and $$mB = -j\omega LI_g$$

$$B = -\frac{j\omega LI_g}{m} = -\frac{j\omega LI_g}{\omega\sqrt{LC}}$$

$$= -j\sqrt{\frac{L}{C}}\,I_g$$

Hence equations become

$$E_x = E_g \cos mx - j\sqrt{\frac{L}{C}}\,I_g \sin mx \quad . \quad . \quad (10)$$

$$I_x = I_g \cos mx - j\sqrt{\frac{C}{L}}\,E_g \sin mx \quad . \quad . \quad (11)$$

We cannot eliminate the unknown I_g unless we know the terminating conditions of the line.

Suppose the line to be infinitely long, then no reflection is possible. Let Z_0 then be the effective impedance of the line from the generator end. Then $I_g = \frac{E_g}{Z_0}$. The impedance of the line at any finite distance x from the generator must still be Z_0, because there is still an infinite length of the line lying beyond the point considered. In order to find Z_0 take a value of x such that $mx = \frac{\pi}{2}$ so that $\cos mx = 0$ and $\sin mx = 1$.

Then $$\frac{E_{x1}}{I_{x1}} = Z_0 \quad \frac{-j\sqrt{\frac{L}{C}}\cdot\frac{E_g}{Z_0}}{-j\sqrt{\frac{C}{L}}\cdot E_g}$$

or
$$Z_0^2 = \frac{L}{C}$$

$$Z_0 = \sqrt{\frac{L}{C}} \quad . \quad . \quad . \quad . \quad (12)$$

Hence the infinite line loads the generator in exactly the same manner as a pure resistance of value $\sqrt{L/C}$. This value may be denoted by Z_0 and is termed the characteristic impedance of the line.

It is important to notice that although the value of Z_0 in ohms is given by $\sqrt{L/C}$ (where L and C are expressed in henrys and farads respectively), yet Z_0 is an effective resistance, not a reactance. We have also shown that the impedance at any point along the infinite line is Z_0 and therefore the current is in phase with the voltage and of value $\frac{E_x}{Z_0}$. Suppose now we have a finite length of line, then if we place a resistance equal to Z_0 across the far end there will be no reflection because we have placed there a resistance equivalent in every way to the impedance of the line to infinity. From another point of view we have inserted a termination which behaves in the same way to the wave as the line behaves, and therefore there is no reflection because there is no change of medium. Evidently this is the correct termination to use for most efficient transmission of energy, since no energy will now be wasted in a system of stationary waves superimposed on the travelling wave. Notice also that since the terminating resistance Z_0 is replacing the portion of line cut off, which extends always to infinity, the termination is correct anywhere; and hence the actual length of line used is of no account (except that as lines are not loss-less the longer the line used the greater the losses involved) and the correct termination will always be Z_0. Because the line has actually resistance and insulation loss, the correct termination will be a complex impedance, to determine which the full analysis of the transmission line as used by the telephone engineer would have to be applied, but the difference in the case of radio-frequency lines is small.

The equations for the voltage and current at any point in

an infinite or correctly terminated line may, therefore, be written

$$E_x = E_o\,(\cos \omega\sqrt{LC}\,x - j \sin \omega\sqrt{LC}.x) \quad . \quad (13)$$

$$I_x = \frac{E_o}{Z_o}\,(\cos \omega\sqrt{LC}\,x - j \sin \omega\sqrt{LC}.x) \quad . \quad (14)$$

These expressions will evidently repeat themselves every time $\omega\sqrt{LC}.x$ is increased by 2π or when x is increased by $\frac{2\pi}{\omega\sqrt{LC}}$. In other words the wavelength $\lambda = \frac{2\pi}{\omega\sqrt{LC}}$ (15)

For any wave motion $v = f\lambda$

$$\therefore \qquad v = \frac{\omega}{2\pi} \cdot \frac{2\pi}{\omega\sqrt{LC}} = \frac{1}{\sqrt{LC}} \quad . \quad . \quad . \quad (16)$$

This velocity is equal to that of a "free" electromagnetic wave in a pure dielectric, the conductors having been assumed to be mere guides, in which no loss is taking place. If the dielectric is air then v will be 3×10^8 metres per second. This can be confirmed by substituting the values of L and C for parallel lines, for example, as given on page 173. In all practical cases, the permeability of the medium between the conductors will be approximately unity but may have a dielectric constant greater than that of air, in which case it will be seen that the velocity (in a loss-free line) will be inversely proportional to the square root of the dielectric constant.

It is convenient to write θ for $\omega\sqrt{LC}.x$ that is, for $2\pi x/\lambda$. θ may be termed either the "electrical length" or the "angular length" of the line. Then (13) and (14) become

$$E_\theta = E_o\,(\cos \theta - j \sin \theta) \quad . \quad . \quad . \quad (17)$$

$$I_\theta = \frac{E_o}{Z_o}\,(\cos \theta - j \sin \theta) \quad . \quad . \quad . \quad (18)$$

It will be seen that the velocity of waves of all frequencies is the same along the loss-free line. In the case of actual lines this is a good approximation for all radio frequencies though it would be far from true, in many cases, at telephone frequencies.

We have thus far considered the line terminated with an actual resistance but except for test purposes such a case would not arise in practice. R.F. lines will normally be terminated by loaded resonant circuits of some form and it is well known

that circuits of this type can act as a resistive load. By the design of suitable coupling circuits, which will be described later, such terminations may be made to match Z_0 exactly.

An interesting relationship exists in a *loss-free line* between the voltage at any point and the current at a point one-quarter wavelength away, in either direction.

If x is increased by $\lambda/4$ then θ is increased by $\pi/2$ so that

$$E_\theta = E_g \cos\theta - j\, Z_0 I_g \sin\theta$$

$$I_{\theta+\frac{\pi}{2}} = I_g \cos\left(\theta + \frac{\pi}{2}\right) - j\frac{E_g}{Z_0}\sin\left(\theta + \frac{\pi}{2}\right)$$

$$= -\, I_g \sin\theta - j\frac{E_g}{Z_0}\cos\theta$$

$$= -\, j\frac{E_\theta}{Z_0} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (19)$$

Hence, whatever may be the terminating conditions in a loss-free line, the voltage at any point is equal in magnitude to the current at a point $\lambda/4$ away, multiplied by Z_0. This relationship has been used to estimate high voltages on a line by current measurements, which are easier to make at high frequencies.

Characteristic Impedance of Parallel Wires

The value of Z_0 can evidently be calculated in cases where the dielectric constant of the medium between the wires is known.

The capacitance per metre between two parallel wires far removed from other conductors is (in rationalised M.K.S. units)

$$\frac{\pi\kappa}{\log_e \dfrac{d}{r}} \text{ farads,}$$

where d = distance between centres of wires, r = radius of wires (both in metres).

If dielectric is air $\left(\kappa_0 = \dfrac{10^{-9}}{36\pi}\right)$ this becomes

$$\frac{10^{-8}}{828 \log_{10} \dfrac{d}{r}} \text{ farads per metre.}$$

The inductance per metre of two wires forming a loop is $\frac{\mu}{\pi} \log_e \frac{d}{r}$ and if μ is μ_o, that is $4\pi \times 10^{-7}$, this becomes

$$9{\cdot}21 \times 10^{-7} \log_{10} \frac{d}{r} \text{ henrys per metre.}$$

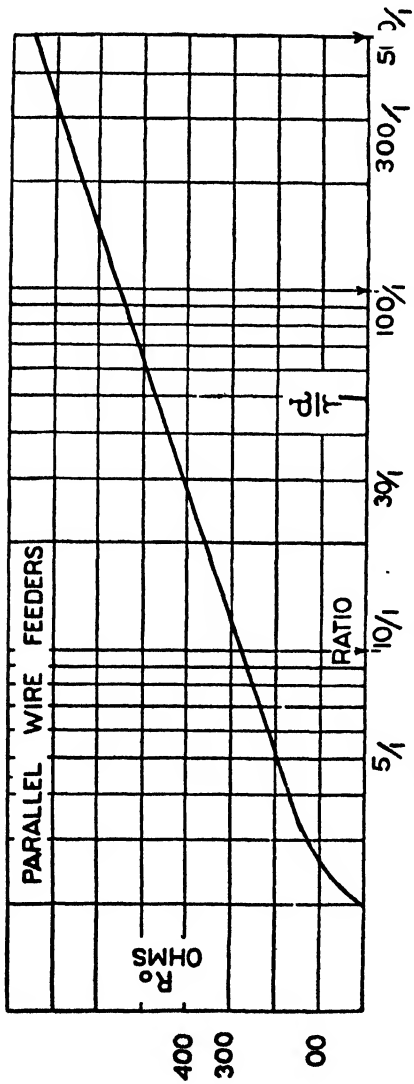


FIG. 88. Variation of Z_0 with d/r—Parallel-wire Line.

Hence $$Z_0 = \sqrt{\frac{L}{C}} = 276 \log_{10} \frac{d}{r} \text{ ohms} \quad . \qquad (20)$$

This formula is accurate down to a ratio of 4/1 for d/r and Fig. 88 shows a curve of Z_0 plotted against d/r obtained by using the above formula but corrected for ratios below 4/1. It will be seen that usual spacings and wire diameters give values of Z_0 in the neighbourhood of 500 ohms.

Lines consisting of four parallel wires are also sometimes

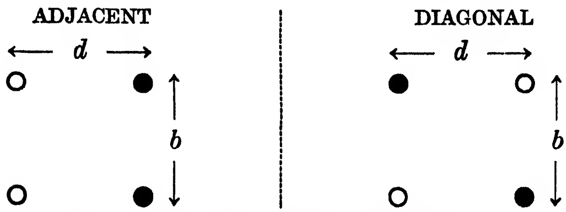


FIG. 89. Alternative Types of Four-Wire Line.

used. The wires may be run either with the go and return wires adjacent or diagonal, as shown in Fig. 89.

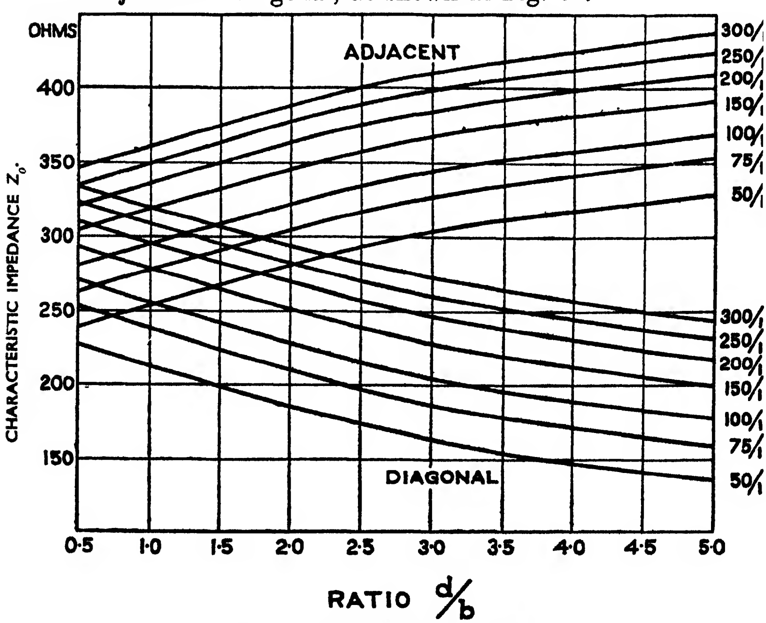


FIG. 90. Z_0 for Four-Wire Lines.

ADJACENT		DIAGONAL	
$Z_o = 138 \log_{10} \dfrac{d\sqrt{d^2+b^2}}{br}$	(21)	$Z_o = 138 \log_{10} \dfrac{db}{r\sqrt{b^2+d^2}}$	(22)
When $d = b$		When $d = b$	
$Z_o = 138 \log_{10} \sqrt{2}\dfrac{d}{r}$.	(23)	$Z_o = 138 \log_{10} \dfrac{d}{r\sqrt{2}}$.	(24)

The curves for four-wire lines are shown in Fig. 90, from which it is observed that both have a smaller characteristic impedance than the twin wire; that the values for Z_o approach equality when the ratio d/b is very small; that the curves for the adjacent spacing rise to become asymptotic to values for the twin wire when the spacing d/b is large; and that an increase of the ratio d/b decreases the characteristic impedance of the diagonally-spaced line.

In general this will mean that for diagonal spacing it will be desirable to keep up the ratio d/b whereas for adjacent spacing it should be kept down, but in both cases a value of d/b of unity is a good starting point for consideration.

Characteristic Impedance of Concentric Tube or Co-axial Lines

In this case the capacitance per metre is

$$\frac{2\pi\ \kappa}{\log_e \dfrac{r_2}{r_1}} \text{ farads,}$$

where r_2 is the inner radius of the outer tube and r_1 the outer radius of the inner tube. When κ is κ_o this becomes

$$\frac{10^{-9}}{41{\cdot}5 \log_{10} \dfrac{r_2}{r_1}} \text{ farads,}$$

Inductance per metre is $\dfrac{\mu}{2\pi} \log_e \dfrac{r_2}{r_1}$

or $4{\cdot}61 \times 10^{-7} \log_{10} \dfrac{r_2}{r_1}$ henrys per metre if $\mu = \mu_o$.

Hence $$Z_o = 138 \log_{10} \frac{r_2}{r_1} \text{ ohms} \qquad (25)$$

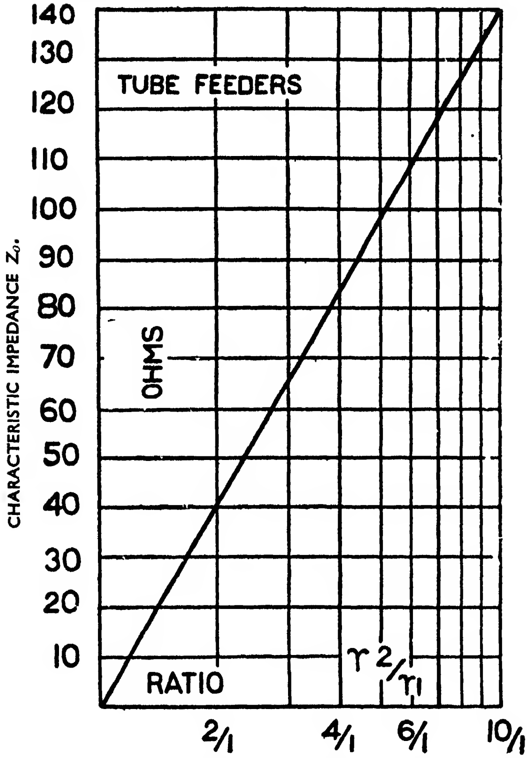


FIG. 91. Z_o for Concentric Lines.

Usual ratios of r_2/r_1 give values of about 100 ohms for Z_o.

We will now study the effects when open or shorted lines of finite length are connected across a source of sinusoidal voltage.

Open-circuited Line

Suppose now that we have a line of electrical length θ_1 which is open-circuited at the far end. Then voltage and current equations at θ_1 become

$$E_{\theta 1} = E_g \cos \theta_1 - jZ_o \sin \theta_1$$

$$0 = I_g \cos \theta_1 - j \frac{E_g}{Z_o} \sin \theta_1$$

From (19) $$I_g \cos \theta_1 = j \frac{E_g}{Z_o} \sin \theta_1$$

$$I_g = j\,\frac{E_g}{Z_o}\tan\theta_1$$

The impedance of the line from the generator end is therefore

$$Z_{oc} = -jZ_o\cot\theta_1 \quad . \quad . \quad . \quad . \quad (26)$$

Substituting the value obtained for I_g in the general equations (10) and (11), we have

$$E_\theta = E_g(\cos\theta + \tan\theta_1\sin\theta)$$

$$I_\theta = j\,\frac{E_g}{Z_o}(\tan\theta_1\cos\theta - \sin\theta)$$

By multiplying and dividing these equations by $\cos\theta_1$ they become

$$E_\theta = E_g\,\frac{\cos(\theta_1-\theta)}{\cos\theta_1} \quad . \quad . \quad . \quad . \quad (27)$$

$$I_\theta = -\frac{jE_g\sin(\theta_1-\theta)}{Z_o\cos\theta_1} \quad . \quad . \quad . \quad (28)$$

An examination of these equations shows that the voltage has the same phase at all points along the line, that is, we have a pure stationary wave ; the current is in quadrature with the voltage ; there are a series of positions along the line, $\lambda/2$ apart, at which the voltage is always zero (voltage nodes) and, displaced $\lambda/4$ from the voltage nodes, a set of current nodes. Further, it will be seen that certain values $(\frac{\pi}{2}, \frac{3\pi}{2}, \ldots)$ of θ_1, that is, certain lengths of line, are favourable to the formation of large stationary waves. In fact, our approximate theory suggests that the voltages and currents will reach infinity at certain points. The actual values reached will depend upon the losses, just as the current in a series resonant circuit depends only upon the resistance.

Considering further the input impedance of the open-circuited line, it will be seen from Fig. 92 that when θ_1 is less than $\frac{\pi}{2}$ the line is equivalent to a capacitance and that when θ_1 is $\frac{\pi}{2}$, that is, the line is $\lambda/4$ long, then its input impedance is zero. After experience at low frequencies, it seems strange to open-circuit one end of a short length of line and yet have

the effect of a short-circuit at the other. When the line exceeds $\lambda/4$ then cot θ_1 becomes negative and, therefore, the input impedance becomes an inductive reactance. The variation of input impedance with length of line is shown in Fig. 92.

It will be seen that an open-circuited line $\lambda/4$ long is in some ways analogous to a series resonant circuit. The voltage across part of the circuit (the open end) may be many times greater than that applied from the generator. The current

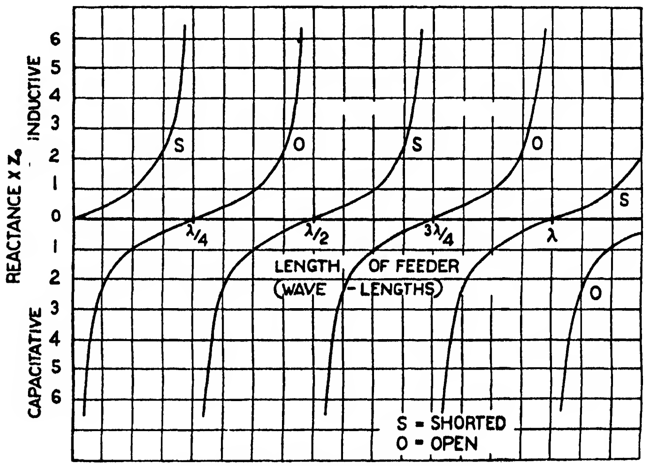


FIG. 92. Reactance of Lines.

from the generator is large and determined entirely by the losses, the load on the generator being purely resistive. An important difference is that the resonant circuit tunes to only one frequency but if the frequency applied to the line is increased so that it is $3\lambda/4$ long, then conditions are the same as for $\lambda/4$.

Short-circuited Line

If the end of a line of electrical length is closed by a perfect short circuit, then at θ_1

$$0 = E_g \cos \theta_1 - jZ_o I_g \sin \theta_1$$

and

$$I_{\theta 1} = I_g \cos \theta_1 \quad \cdot j \frac{E_g}{Z_o} \sin \theta_1$$

$$I_g = j\frac{E_g}{Z_o}\cot\theta_1$$

The reactance of the line, considered from the generator end, will therefore be

$$Z_{sc} = jZ_o\tan\theta_1 \quad . \quad . \quad . \quad . \quad (29)$$

Hence when the line is less than $\lambda/4$ it will have an input impedance which is a pure inductive reactance, whilst a $\lambda/4$ line will have an infinitely great impedance. The variation of input impedance with length for a short-circuited line is also shown in Fig. 92, the curves being of the same shape as those for the open-circuited line but shifted through $\lambda/4$.

Substituting for I_g in the general equation and rearranging, we have

$$E_\theta = \frac{E_g\sin(\theta_1-\theta)}{\sin\theta_1} \quad . \quad . \quad . \quad (30)$$

$$I_\theta = -\frac{jE_g}{Z_o}\frac{\cos(\theta_1-\theta)}{\sin\theta_1} \quad . \quad . \quad . \quad (31)$$

As before, we have a pure stationary wave. The $\lambda/4$ line is seen to be in many ways analogous to a parallel resonant circuit, since it presents an infinite impedance across the generator or, if losses are allowed for, a very high resistance which increases as losses decrease. The generator voltage is the biggest voltage present but the current in the rest of the circuit is larger than that of the generator.

Line Terminated by any Impedance Z_r

If the line is of electrical length θ_1 then

$$E_{\theta 1} = E_g\cos\theta_1 - jZ_oI_g\sin\theta_1$$

$$I_{\theta 1} = I_g\cos\theta_1 - j\frac{E_g}{Z_o}\sin\theta_1$$

But if the line is terminated by an impedance Z_r, then $E_{\theta 1} = I_{\theta 1}Z_r$.

Hence $E_g\cos\theta_1 - jZ_oI_g\sin\theta_1 = I_gZ_r\cos\theta_1 - jE_g\frac{Z_r}{Z_o}\sin\theta_1$

or
$$I_g = E_g\frac{\cos\theta_1 + j\frac{Z_r}{Z_o}\sin\theta_1}{Z_r\cos\theta_1 + jZ_o\sin\theta_1}.$$

Substituting this value for I_g in the general equation (10) and rearranging gives

$$E_\theta = E_g \frac{Z_r \cos(\theta_1 - \theta) + jZ_o \sin(\theta_1 - \theta)}{Z_r \cos\theta_1 + jZ_o \sin\theta_1} \qquad (32)$$

$$I_\theta = \frac{E_g}{Z_o} \frac{Z_o \cos(\theta_1 - \theta) + jZ_r \sin(\theta_1 - \theta)}{Z_r \cos\theta_1 + jZ_o \sin\theta_1} \qquad (33)$$

The sending-end impedance

$$Z_g = E_g/I_g = Z_o \frac{Z_r \cos\theta_1 + jZ_o \sin\theta_1}{Z_o \cos\theta_1 + jZ_r \sin\theta_1} \qquad (34)$$

Properties of a Quarter-wave Line

We have already seen that a $\lambda/4$ line open-circuited at the far end has zero input impedance, whilst a short-circuited line has infinite input impedance. These are the extremes of a general condition for if one end of a $\lambda/4$ line is terminated by a pure resistance Z_o/n then the input impedance is a resistance of value nZ_o. This may be proved as follows :

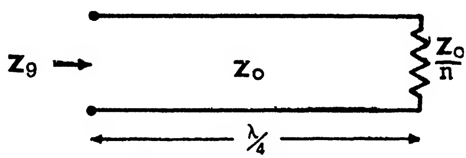


FIG. 93. Quarter-Wave Line.

For a line $\lambda/4$ long $\left(\theta = \frac{\pi}{2}\right)$, (34) becomes $Z_g = Z_o^2/Z_r$.

Hence, if $Z_r = \frac{1}{n} Z_o$, $\qquad Z_g = nZ_o$.

A similar argument will show that a terminating inductive reactance of value Z_o/n appears at the generator end of a $\lambda/4$ line as a capacitative reactance of nZ_o.

A $\lambda/4$ line can therefore be used as a transformer for matching low impedances to high ones, or vice versa. A common application of such a transformer is to terminate a feeder of characteristic impedance Z_o supplying an aerial equivalent to a resistance R, as shown in Fig. 94. A $\lambda/4$ length of line having the correct characteristic impedance Z_{o2} will terminate the main line correctly.

There will be stationary waves set up in the $\lambda/4$ line, because it is not correctly terminated, but the conductor and insulator

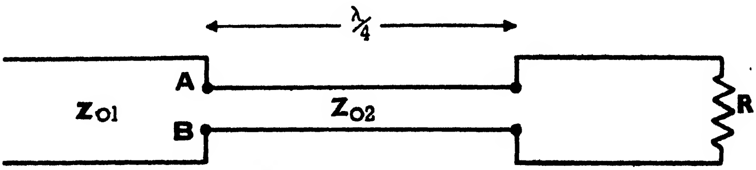


FIG. 94. Matching by Quarter-Wave Line.

losses thereby introduced are not serious because of the short length concerned.

From $$Z_{AB} = \frac{Z_{o2}^2}{R} \text{ and } Z_{AB} \text{ is to equal } Z_{o1}$$

$$\therefore \quad Z_{o2} = \sqrt{Z_{o1} R} \qquad (35)$$

Problem (1). A half-wave aerial, tuned to 20 megacycles, is equivalent (between its feedpoints) to a non-inductive resistance of 100 ohms. It is to be fed by a parallel-wire feeder having a characteristic impedance of 500 ohms. Design a suitable transforming feeder to insert between the main feeder and the aerial so that the former may be correctly terminated.

Solution (1). The feeder should be $\lambda/4$ long. If the velocity along it may be taken to be the same as in free space, then the required length is 3·75 metres. (Actually, the velocity will be somewhat less due to losses and to the presence of supporting insulators, and the correct length will therefore be reduced. A usual value of velocity for parallel-wire feeders supported at intervals by good insulators is some 90% of that in free space.)

The characteristic impedance of the transforming feeder should be $\sqrt{500 \times 100} = 224$ ohms. Reference to Fig. 88 shows that this could be realised by using two rods of 0·5 cm. radius, spaced 3·2 cm apart.

Equivalent of Series and Parallel Impedances

A general theorem may here be stated, as it is frequently useful in connection with line calculations. Any impedance can evidently be expressed either as a resistance and a reactance in series or a different resistance and reactance in parallel. Thus consider the circuits (*a*) and (*b*) of Fig. 95.

If (*a*) represents an actual arrangement of condenser and

resistance in series it is nevertheless possible to find values of a condenser and resistance which placed in parallel would take the same current as the series circuit when the same voltage at the same frequency was applied to it.

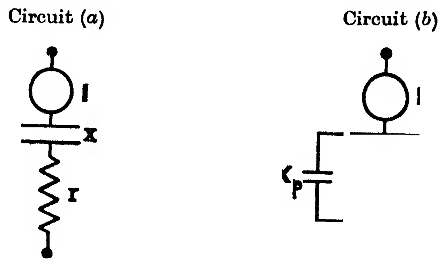


FIG. 95. Equivalent Series and Parallel Impedances.

$$I = \frac{E}{r - jx} = \frac{E\,(r + jx)}{r^2 + x^2} \qquad I = \frac{E}{R_p} + \frac{E}{-jX_p}$$

Hence $$\frac{r}{r^2 + x^2} = \frac{1}{R_p} \text{ and } \frac{x}{r^2 + x^2} \quad \frac{1}{X_p}$$

or $$R_p = \frac{r^2 + x^2}{r} \text{ and } X_p = \frac{r^2 + x^2}{x} \qquad . \quad (36)$$

Conversely the conversion of parallel elements of X_p and R_p into equivalent series elements r and x, can be carried out by the following formula.

$$r = R_p \frac{X_p^2}{R_p^2 + X_p^2} \qquad . \quad (37)$$

$$x = X_p \frac{R_p^2}{R_p^2 + X_p^2} \qquad . \quad (38)$$

or if r has previously been found x can be obtained more simply from

$$x = \frac{R_p r}{X_p} \qquad . \quad (39)$$

By the use of this theorem, together with our knowledge of the properties of a $\lambda/4$ line, we can deduce the values of r and x

by measuring the impedance between A and B as a parallel combination R and X, Fig. 96.

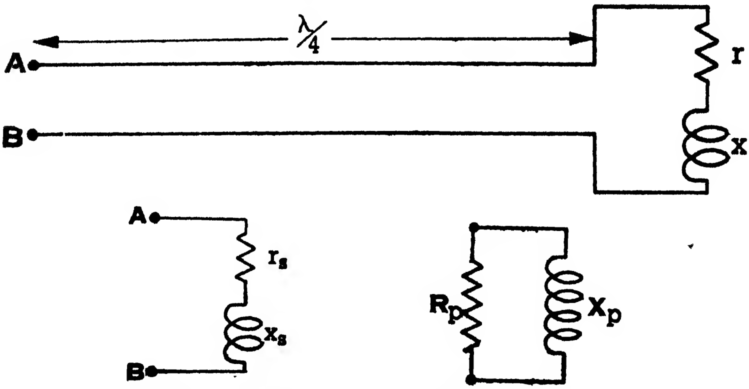


Fig. 96. Impedance Transformation by λ/4 Line.

From (33) $$r_s + jx_s = \frac{Z_o^2}{r + jx}$$

from which $$r_s = \frac{Z_o^2 r}{r^2 + x^2} \text{ and } x_s = \frac{-Z_o^2 x}{r^2 + x^2}$$

From (36) $$R_p = \frac{r_s^2 + x_s^2}{r_s} = \frac{Z_o^2}{r}$$

$$X_p = \frac{r_s^2 + x_s^2}{x_s} = -\frac{Z_o^2}{x}$$

or $$r = \frac{Z_o^2}{R_p} \text{ and } x = -\frac{Z_o^2}{X_p} \quad . \quad . \quad (40)$$

This relationship is sometimes useful when conducting measurements on lines.

The Reactance Transformer

An alternative method for transforming an impedance to another value involves the use of what is known as a reactance transformer. For instance, suppose we have a line with a characteristic impedance Z_o and desire to couple it to a load resistance r whose value is but Z_o/n.

By placing an inductance in series with r we can increase the impedance of this branch but there will then be a large

reactive component. This can be balanced by a parallel capacitance, leaving the net impedance between A and B (Fig. 97) purely resistive and of the correct value.

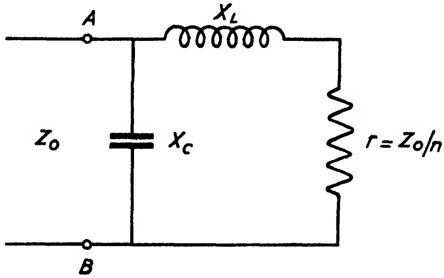


FIG. 97. The Reactance Transformer.

The correct sizes of X_L and X_C can be determined as follows: By the usual rule for parallel impedances,

$$Z_{AB} = Z_o = nr = \frac{(r + jX_L)(-jX_C)}{r + j(X_L - X_C)}$$

$$nr^2 + jnr(X_L - X_C) = X_L X_C - jX_C r$$

Equating the real parts,

$$nr^2 = X_L X_C \text{ or } X_L = \frac{nr^2}{X_C}$$

Equating the unreal parts,

$$nr(X_L - X_C) = -X_C r$$

$$n\frac{nr^2}{X_C} - nX_C = -X_C$$

$$X_C = \frac{nr}{\sqrt{n-1}} \quad . \quad . \quad . \quad . \quad (41)$$

and

$$X_L = r\sqrt{n-1} \quad . \quad . \quad . \quad . \quad (42)$$

It is not necessary that the load impedance should be simply a pure resistance r. If the load has actually inductive reactance for example, then X_L may be reduced to give the same total inductive reactance for the branch.

The arrangement can, of course, be used in the reverse direction. Thus the larger resistance R may be load which is to be transformed into a smaller resistance r.

Problem (2). A main feeder branches into two, all the feeders being of the same concentric type with a characteristic impedance of Z_0. It is desired to install a reactance transformer at the junction so that the main feeder may be correctly terminated.

Solution (2). The load impedance of the two feeders in parallel will be $\frac{Z_0}{2}$. Hence $n = 2$ and from our equations we find that $X_C = Z_0$ and $X_L = \frac{Z_0}{2}$. The vector diagram for this simple case is given in Fig. 98.

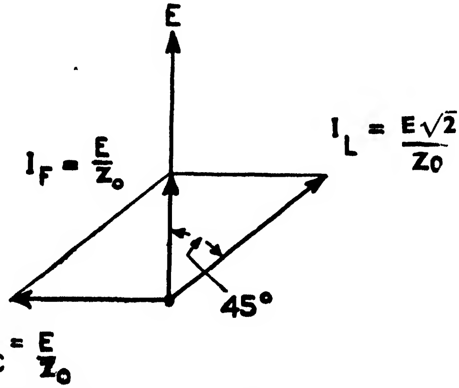


FIG. 98. Vector Diagram for Reactance Transformer.

Problem (3). The base impedance of a given aerial may be considered as a resistance of 300 ohms in series with a capacity reactance of 200 ohms. The feeder used has a characteristic impedance of 75 ohms. Determine the reactances to be used in a suitable transformer.

Solution (3). This is evidently a transformation in the reverse direction to that discussed previously and therefore the connections

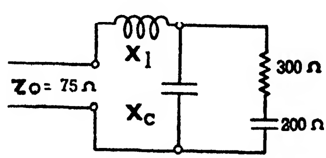


FIG. 99. Illustrating Problem 3.

will be as Fig. 99. It will be necessary to express the aerial impedance in parallel form and from equation (36) these become 433 ohms resistance and 650 ohms capacity reactance. The reactance will

be provided for by increasing the size of X_C and will therefore be ignored at the moment.

$$n = \frac{433}{75} = 5{\cdot}78$$

$$X_C = \frac{433}{\sqrt{4{\cdot}78}} = 198 \text{ ohms and } X_L = 75\sqrt{4{\cdot}78} = 164 \text{ ohms.}$$

It will be seen that X_C is made up of two parallel reactances, 650 ohms due to the aerial and X say, due to the condenser in the transformer. We therefore have $\frac{1}{650} + \frac{1}{X} = \frac{1}{198}$ from which $X = 286$ ohms.

Graphical Solution of Line Problems

Graphical solutions are useful in many engineering problems because they frequently allow an approximate solution to be obtained more quickly than by calculation, with less risk of gross error. Also, such methods usually give a good insight into the physical nature of the problem. When studying the behaviour of the R.F. line, especially if it is assumed free from loss, graphical methods are simple and effective and link up directly with the measurements usually made in practical problems.

In our analytical study of lines, we commenced with a known voltage applied to the line and measured everything from this. In the graphical treatment, since the load end determines the relative magnitudes and phases of the various currents and voltages (which are often all we require) we consider this end first and measure back from it.

Consider a wave started from the generator end. At every point the current will be given by the voltage divided by the characteristic impedance. Unless the termination equals Z_0, there will be a reflected wave formed (in which, also, the current is given by the voltage divided by Z_0) and unless the generator matches the line this reflected wave will give rise to another forward wave superimposed on the first and so on. The final distribution of current and voltage along the line is therefore the sum of innumerable components but since all the forward waves are of the same frequency and travel in the same direction (although not necessarily of the same time phase), they can be compounded into one incident wave and similarly for the

reflected waves. Hence the actual current and voltage at any point can be considered as the resultant of one incident wave whose current I_f is in time phase with the voltage V_f, and one reflected wave whose current I_r is in time phase with the reflected voltage V_r.

Where the terminating circuit contains no power dissipating element (any pure reactance or an open or short-circuit) the reflected wave will be equal in magnitude to the incident wave and a series of stationary waves with zero current minima will be formed as shown in Fig. 100a, voltage and current waves being in quadrature both as regards space and time. If, however, the circuit has a resistive element (not equal to Z_0) only part of the

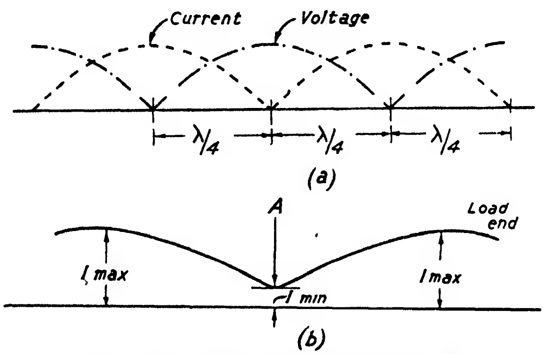


FIG. 100. Stationary Waves on Lines.

incident wave will be absorbed by the termination, the rest being reflected. We can then imagine a "wattful" current (and voltage) wave travelling forward to supply the load power, on top of which rides a stationary current (and voltage) wave, the current wave being shown in Fig. 100b. It will be shown that there is a definite relationship between the minimum/maximum ratio and the circuit resistance in terms of Z_0 and that a simple measurement of the ratio in purely relative terms can be used to solve many line problems.

The type of stationary wave shown in Fig. 100a is typical of a line whether open-circuited, shorted, or terminated by a pure reactance, the only changes being in the space position of the wave and in its magnitude. With an open-circuited line there will obviously be zero current at the end, and the first maximum

will occur $\lambda/4$ back (or 90°); with a shorted line the reverse will hold. With a pure reactance the position will depend upon the value of the reactance in terms of Z_0 but in no case, of course, will either a maximum or minimum occur at the end.

The stationary wave shown in Fig. 100b is characteristic of a line terminated by pure resistance, or by any complex impedance. If the terminating circuit is resistive and of value less than Z_0, a current maximum occurs at the end (the lower the resistance the greater the terminal current) and for resistances greater than Z_0, a current minimum will occur at the end, approximating to zero with resistance values very much

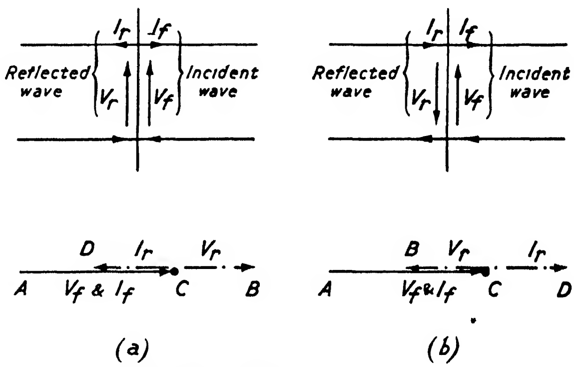


FIG. 101. Vector Convention.

greater than Z_0. With any complex impedance the wave will assume some position between the two just mentioned.

The graphical solutions involve the setting up of a vector diagram so that at the termination, current and voltage vectors are built up from the vector sum of incident and reflected currents and voltages. The vectors must be able to give space conditions along the line, as well as time changes. We have already seen that $I_f = V_f/Z_0$ and $I_r = V_r/Z_0$. If the line is free from loss, then Z_0 is a pure resistance and, therefore, I_f is in phase with V_f and I_r in phase with V_r. Also I_f and V_f (and I_r and V_r) are constant in magnitude at all points along the line although I_r may be less than I_f and V_r less than V_f. Since $I_fZ_0 = V_f$ and they are always in time phase we can use two coincident vectors of equal length to show them. Similarly, $I_rZ_0 = V_r$ and they are also always in time phase.

Thus they can be represented by two vectors of equal magnitude, not coincident but 180° apart, because, for a charge of a given polarity on the line, the reflected current is in opposite direction to the incident current, and any convention of vectors must show this. For instance, if at the point of a line considered, as Fig. 101a, (top) incident and reflected voltages have the same polarity and the reflected voltage is half the incident, the vector representation of both waves will be as shown in Fig. 101a, where AC represents both incident voltage V_f, and current $I_f Z_0$, CB shows the reflected voltage V_r, and CD the reflected current $I_r Z_0$. The voltages add to give a voltage of $V = AB$, and the

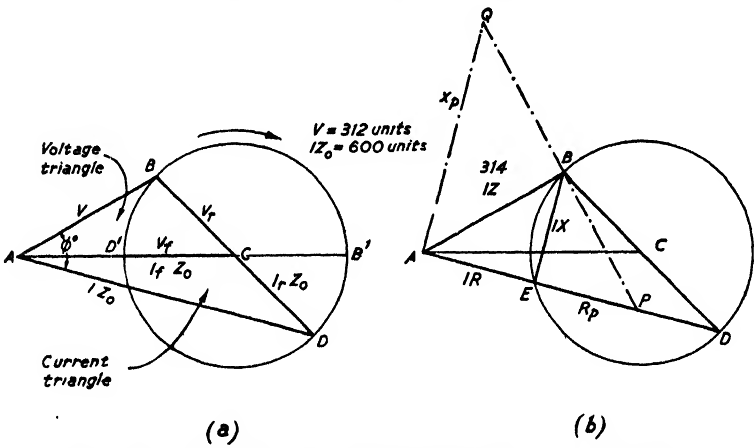


FIG. 102. Vector Diagrams for Incident and Reflected Waves.

currents subtract to give a current at that point of $AD = IZ_0$. Had the waves passing been such that the polarity of incident and reflected voltages was opposite as shown in Fig. 101b (top) the currents are now in the same direction in the lines. Vectorially these currents and voltages are shown below the corresponding figure from which it is seen that the vector DCB is merely reversed in direction so that voltages subtract and currents add.

At most points along the line the component waves will not be in phase as discussed above but will have a phase angle between them. The vectors will then open out into two triangles as shown in Fig. 102a where the various components have the same significance as before, namely AC

is the forward wave vector, BC, CD the reflected wave vector pair, and AB and AD are the resultant voltage and current. It will be observed that we have a voltage triangle ACB and a current triangle ACD, each component in the latter being multiplied by the common numerical factor Z_0. Rotation of these triangles anti-clockwise about C will show the cyclic changes of currents and voltages by projection of the vectors to the vertical axis in the usual way at the point considered.

Let us consider a general case, say of a line of characteristic impedance Z_0 terminated by an inductive impedance $Z\,\underline{|\phi}$. Since $Z\,\underline{|\phi}$ is known we can set up vectors V and I, $\phi°$ apart, remembering that I is multiplied by the factor Z_0. But since V is the vector sum of V_f and V_r and I the sum of I_f and I_r, we complete the triangles by joining B to D, bisecting BD in C, since $V_r = I_v Z_0$, and joining AC which give the triangles we have shown in Fig. 102a.

Problem (4). Find the terminal conditions for a 600 ohm line terminated by an inductive impedance of 314 ohms, phase angle 43°.

Solution (4). If 1 amp. is assumed to flow through the impedance, then :—

$$V = IZ = 314 \text{ volts}, \; I = 1 \text{ amp.}, \; \phi = 43° \text{ lagging.}$$

These values are shown in Fig. 102b, where $AB = 314$ volts $= 314$ units, and $AD = 1$ amp., made 600 units long since I is multiplied by Z_0, the vectors being 43° apart, V leading.

Completing the construction as indicated above and scaling voltages and currents in their appropriate units we have :—

V_f	Incident voltage	$= AC$	$= 430$	Voltage triangle.
V_r	Reflected „	$= CB$	$= 215$	
V	Voltage	$= AB$	$= 314$	
I_f	Incident current	$= AC$	$= \cdot 72$	Current triangle.
I_r	Reflected „	$= CD$	$= \cdot 36$	
I	Current	$= AD$	$= 1 \cdot 00$	

A complex impedance such as we have discussed as a termination may represent a voltage-current condition along a line and we may require to evaluate the series or parallel components of resistance and reactance, usually the parallel. This may be done graphically by the simple additional construction shown in Fig. 102b which gives the values for the example just analysed.

Problem (5). Evaluate the series X and R, and parallel X_p, R_p for the impedance case given above.

Solution (5). From the previous figure (Fig. 102b) drop a perpendicular from the point B to the line AD, giving the point E. This falls within AD (as shown). It would fall outside the point D when $R > Z_0$. Then the triangle formed by ABE gives the series values required, namely $AB = V = ZI$, $BE = XI$ and $AE = RI$. The values obtained by scaling ABE are :—

$$\begin{aligned} AB &= Z = 314 \text{ ohms} \\ BE &= X = 221 \quad " \\ AE &= R = 234 \quad " \end{aligned}$$

The parallel components R_p and X_p can, of course, be obtained from the above from

$$R_p = R + \frac{X^2}{R} = 442\Omega \text{ and } X_p = X + \frac{R^2}{X} = 469\Omega$$

or graphically by drawing the line QP at right angles to the vector AB at B, to cut the line AD in P (or P may fall outside D) and erecting a perpendicular from A to meet Q. Then AQ is equal to X_p and $AP = R_p$, the scale for these values still being AB. Values obtained from this figure give

$$\begin{aligned} AB &= Z = 314 \text{ ohms} \\ AP &= R_p = 442 \quad " \\ AQ &= X_p = 469 \quad " \end{aligned}$$

In the above example we were given a complex impedance and phase angle. Had we been given the component values of R and X we should first have to construct the component triangle ABE before proceeding.

Thus a line AE would first be drawn horizontal of length R (I) to give R, then EB would be set up at right angles of length X (I), upwards if inductive or downwards if capacitive, and B joined to A to give the value of Z (I). The vector would then be completed as previously indicated it being remarked that if R is of greater value than Z_0, the point D will fall inside the line AE and not outside as shown in Fig. 102b. This will not affect the construction as D will be joined to B and its bisection at C will determine the voltage and current diagrams.

Returning to the general problem (Fig. 102a), let us now see how V and I can be determined at a distance $x = \lambda\theta/2\pi$ back from the load. Since we are now " meeting " the incident wave V_f will be advanced in phase by θ, anti-clockwise vectorially, whilst V_r will be retarded by θ, clockwise vectorially

about the point C. The graphical construction, however, will be much simplified if we leave $V_f = AC$ fixed, and retard V_r (and I_r) clockwise by an amount 2θ. Since V_r and I_r are the same length, rotation of these will develop a circle of radius $CB = CD$ indicated in Fig. 102a. Thus if DCB is rotated clockwise through 150° (75° of line length, or nearly $\lambda/5$ back from the load), the point D now falls within AC. We therefore have a current minimum AD' and a voltage maximum AB' and a phase angle of zero. The equivalent input impedance at this point will thus be higher than Z_0 but resistive; and for another 180° of vector rotation, or $\lambda/4$ line length, B' now falls within AC and the reverse will hold, E will be low, I high and the

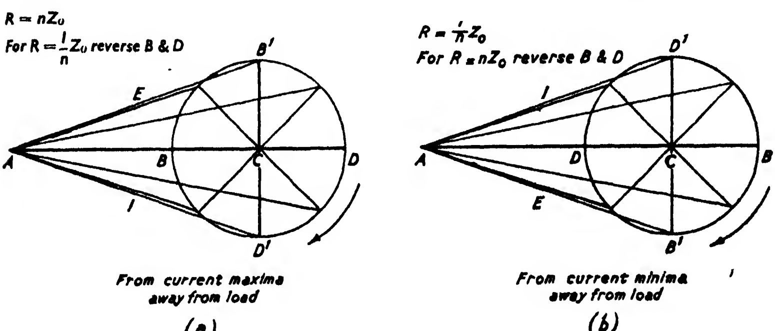


FIG. 103. Vector Diagrams for Resistance Termination.

equivalent impedance will be lower than Z_0 but again resistive. These high and low resistive points are those at which a current measuring instruments records minima and maxima when run along the line as described on page 232. In between these two points mentioned, the impedance is again complex, but sometimes capacitive when D' passes in the half-circle above AC and at other times inductive when D' is in the lower half-circle (shown in Fig. 103a). The various vector conditions are given in Fig. 103 which may be regarded as a universal vector diagram for any termination in terms of Z_0.

Since we have taken a general case we can state that whatever type of termination we have there will be points along the line at which the equivalent impedance is resistive in character.

Let us consider some examples of resistance termination.

Suppose we have a line terminated by a resistance of $R = \frac{1}{2} Z_o$. Then assuming the current magnitude is 1·0 (Z_o), the voltage will be :—

$$\frac{Z_o (V_f - V_r)}{V_f + V_r} = R = \frac{Z_o}{2} \text{ from which } V_r = \tfrac{1}{3} V_f.$$

The vector relationship is shown in Fig. 103a where the vectors have the same significance as formerly. The current is at a maximum $= AD$, and the voltage a minimum

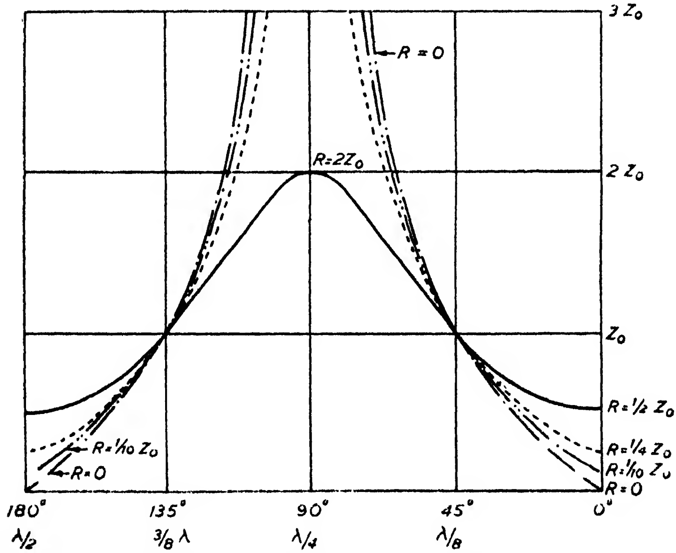


FIG. 104. Variation of Magnitude of Impedance Along a Line Terminated by a Resistance.

AB. If now we consider a point 45° back from the load ($\lambda/8$) by rotating the vector through 90°, V and I have now reached their greatest phase-angle difference (about 36°) V leading and since V and I are now the same length the impedance at this point is equal in magnitude to Z_0 but reactive (inductive since I lags on E). By suitable rotation of the vector BCD we can find the impedance at other points, which have been plotted in Fig. 104 (full line), it being noted that at a point $\lambda/4$ from the load the impedance is now $2Z_0$ and resistive since the vectors are reversed on those initially given, I now

becoming only ·5 amp. and E rising to double. The points of maximum resistance of course coincide with the points of minimum current on the stationary wave and, in the vector analysis, to the vector position when point D lies within AC.

We must discuss this particular condition further. Considering the case just mentioned, let the current-minimum position on the line be at the point A Fig. 100b, which will be given by the vector condition shown by 103b (horizontal vector $ADCB$), it being assumed the load is at the right-hand end of the line as marked. Now if we travel away from the load the condition will be given by clockwise rotation of the vector so that current head of the vector DCB now rises into the upper semicircle. This means that since the current leads, the equivalent impedance for the next $\lambda/4$ of line must be capacitive towards the generator. Conversely if we travel from a current minima towards the load, since the current head of the vector DCB passes into the lower semicircle, the immediate $\lambda/4$ of line must have an inductive impedance. The reverse conditions hold from a current maximum.

From Fig. 103 we have

In magnitude $\qquad I_{min} Z_0 = V_f - V_r$

and $\qquad I_{max} Z_0 = V_f + V_r$

Hence $\qquad \dfrac{I_{min}}{I_{max}} = k = \dfrac{V_f - V_r}{V_f + V_r}$ from which $V_r = \dfrac{1-k}{1+k} V_f$. (43)

Let us consider further the resistance termination of value nZ_0, when n is a fraction. Then a current maximum will occur at the load and the load current and voltage will be given by

$$I_{max} Z_0 = V_f + V_r$$

$$V_{min} = V_f - V_r$$

But $\qquad V/I = nZ_0$

and hence $\qquad \dfrac{V_f - V_r}{V_f + V_r} = n$ . . . . (44)

that is $n = k$, so that the ratio of minimum to maximum stationary wave current gives directly the fraction that the terminating resistance is of Z_0.

If the terminating resistance is larger than Z_0 so that $n > 1$

then a current minimum and a voltage maximum occur at the load and we have

$$\frac{V_f + V_r}{V_f - V_r} = n = \frac{1}{k} \tag{45}$$

In setting up the vectors for any given termination we can make V_f any convenient length, since its actual value is unknown, and the radius of the reflection circle determined by V_r can then be drawn directly from the above. Thus in the case we have just given of $R = \frac{1}{2} Z_0$, $k = \frac{1}{2}$ and the radius V_r is :

$$V_r = \frac{1 - k}{1 + k} V_f,$$

namely $\frac{1}{3}$, a result we have previously obtained.

The impedance changes for different values of R are shown in Fig. 104, it being observed that all the curves pass through Z_o at $\lambda/8$ from the current minima or maxima, although the

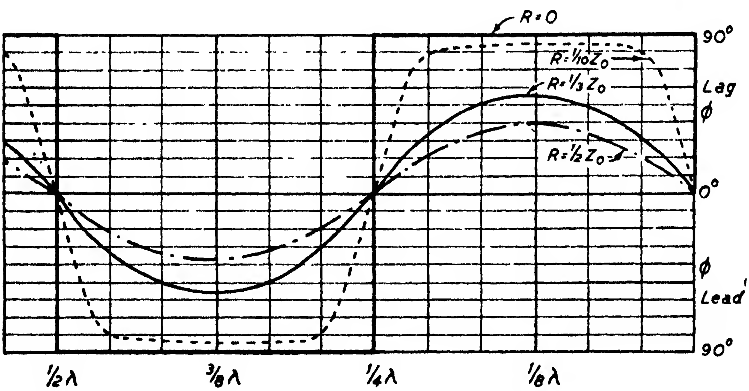


FIG. 105. Variation of Phase of Impedance Along a Line Terminated by Resistance.

phase angles are, of course, different, as shown by Fig. 105, it being noted that the phase changes abruptly to 90°, when the termination is either a short-circuit, open-circuit, or a pure reactance.

Vectorially, the various conditions along the line from any current maxima or current minima are best seen by considering the vector diagrams shown in Fig. 103, space conditions for

each $22\frac{1}{2}°$ of line away from load being given. The radius of the reflection circle will be determined from the Table, which also gives the ratio Z/Z_0 and $I_{\min}/I_{\max}$. Consideration of this diagram and the figures in the Table indicate that with R nearly equal to Z_0 the reflection circle is very small compared with the vector AC. As R diverges more and more from Z_0, the

TABLE VIII

$AC = V_f = 1{\cdot}0$

R in terms of Z_0	0 or ∞	$\frac{1}{10}$ or 10	$\frac{1}{5}$ or 5	$\frac{1}{4}$ or 4	$\frac{1}{3}$ or 3	$\frac{1}{2}$ or 2	$\frac{2}{3}$ 1·5
k	∞	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$
V_r in terms of V_f	1·0	·9	·666	·60	·50	·333	·20

When $R > Z_0$, D lies within AC.
" $k < Z_0$, B " " AC.

From current minima D within AC, clockwise rotation for 180° (away from load) is capacitive.

reflection circle gets larger and the circumference approaches nearer to the point A, which rests on it when R is zero or infinity, or the termination is reactive. It will be seen that when $AC = BC$ and becomes a radius, the phase angle is always 90°, changing abruptly from 90° lag to 90° lead, or vice versa, as the point D, or B passes across the vector AC.

Examples will illustrate the further use of this graphical method.

Problem (6). A line having air dielectric and a characteristic impedance of 500Ω has a condenser across the far end, the losses of which are negligible, whilst a voltage at a frequency of 30 Mc/s is applied at the other. The first current minimum, which is zero, occurs at a point 3 metres from the termination. Find the capacitance of the condenser.

Solution (6). Assuming no losses in line or load, with $Z_0 = 500\Omega$ the velocity of propagation is 3×10^8 metres/sec. Hence $\lambda = 10$

metres, and the value of $\theta = 108°$ for min. current. Since there are no losses $V_r = V_f$ and the vectors for the current minimum are as shown in Fig. 106 (horizontal vector ACB). Since we are going

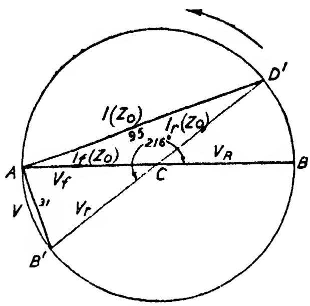


FIG. 106. Illustrating Problem 6.

ON to the load we must advance the vector through 2θ (anti-clockwise), i.e., 216° as shown by the vector $B'CD'$ from which we obtain :—

$$AB' = V = 30$$
$$AD' = I = 95$$

$$\text{Reactance of condenser } \frac{V}{I} \times Z_0 = 158\Omega$$

$$\text{Capacitance of condenser } = 33\mu\mu F.$$

Problem (7). The condenser of the previous example is replaced by an impedance. The first minimum occurs at 3 metres from the load and the ratio $I_{min}/I_{max} = 0{\cdot}2$. Determine the impedance.

Solution (7).

$$V_r = V_f \frac{1 - k}{1 + k} = V_f \frac{1 - 0{\cdot}2}{1 + 0{\cdot}2} = 0{\cdot}667 V_f.$$

The conditions at the current minimum are as shown by the horizontal line (Fig. 107), the current minimum occurring at a point

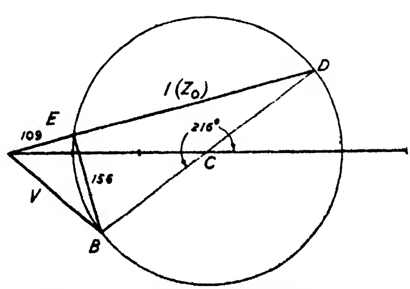


FIG. 107. Illustrating Problem 7.

$\theta = 108°$ from the load. Thus advancing the vector anti-clockwise, as before, 216° we get the vectors given by the triangles ABC, CDA, from which we have :—

$$V = 60,\ I = 158,\ \phi = 55°, \text{ current leading.}$$

$$Z = \frac{V}{I} \cdot Z_0 = 190\Omega \text{ with a leading phase angle of } 55°.$$

We can evaluate the series R and X by calculation or by the graphical construction previously given. Thus : drop from B a perpendicular to AD, giving the point E. Then we have

$$\begin{aligned} AB &= 130 \text{ units} &= 190\Omega &= Z \\ AE &= 91 \quad ,, &= 110\Omega &= R \\ BE &= 156 \quad ,, &= 156\Omega &= X \end{aligned}$$

The load consists then of a resistance of 110Ω in series with a capacity reactance of 156Ω, which is a condenser of $34\mu\mu F$.

Stub Matching

We have seen earlier in this chapter that a line may be matched to a load of different impedance from Z_0 by a reactance transformer, or by a quarter-wave line. Another very useful

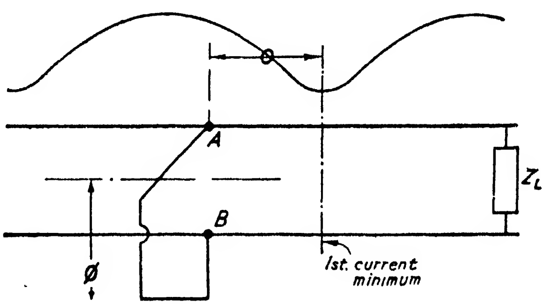


FIG. 108. Stub Matching.

way is by the use of a length of short-circuited, or open-circuited piece of line, known as a stub, as shown diagrammatically in Fig. 108.

We know that when a line is not correctly terminated, the impedance looking towards the load from a current maxima or minima is purely resistive, that from the minima being greater than Z_0 and from the maxima less than Z_0. Between each maxima, or each minima, it is therefore possible to find two points at which the impedance has a series resistance term equal to Z_0 with an unknown X term ; and two other points

at which the impedance can be evaluated in terms of a resistance of R_p equal to Z_0 in parallel with an unknown reactance X_p which may be capacitive or inductive. Actually the series and parallel points are 90° apart in space.

Since the stub is being placed in parallel we are only concerned with the parallel condition. If a shorted stub is used (which will always be less than $\lambda/4$) since it has an inductive reactance, we require the point on the line at which $R_p = Z_0$ and X_p is capacitive. Whereas if the stub is open-circuited, we require a similar point but one where X_p is inductive.

The shape of the stationary wave on a line is a series of half-sine waves (usually smoothed out by the travelling component as shown in Fig. 108, top) which means that the minima can always be picked out much more accurately than the maxima. Thus in practice when taking measurements one always works from a current minimum, usually the first from the load end. If we refer back to Fig. 102a and the discussion on page 195, it was seen that measuring from a current minimum away from the load the reactance is always capacitive, whereas towards the load it is always inductive. This means that a shorted stub will always be placed on the generator side of a current minimum and an open stub on the load side, at points which have to be determined.

We require to find the position on the line from a current minimum at which the parallel value of R_p is equal to Z_0, and the value of X_p with which it is associated. If we refer back to Fig. 102b we see that $AP = IR_p$, where R_p is the parallel resistance component of the impedance. If, therefore, we rotate the vector DCB so that AB becomes tangential to the reflection circle at B, the point P will now coincide with the point D. But $AD = IZ_0 = AP = IR_p$, which is the condition required, and since B is in the upper half-circle, the reactance of X_p given by AQ will be capacitive to which the stub must be matched. The rotation of the vector to give this condition from the current minimum will determine the position of the stub as θ away from the load. Had the vector DCB been rotated in the opposite direction so that the tangent point B is now in the lower half-circle, the conditions are reversed and the line reactance is now inductive. But since BCD has to be rotated either side of the vector AC by the same angle to

give the tangent positions, the points on the line at which $R_p = Z_0$ lie equally on either side of the current minimum.

The length of the stub can be determined graphically; or, since the input impedance of a short-circuited stub, less than $\lambda/4$ is given by $Z_0 \tan \phi$ (see equation 29):

$$X_p = Z_0 \tan \phi \text{ or } \tan \phi = X_p/Z_0.$$

Problem (8). A concentric line with air dielectric, has $Z_0 = 100\Omega$ and is being used at 100 Mc/s. When a load is connected it is found that $k = 0{\cdot}25$ and that the first minimum of current occurs at 1 metre from the load as shown in Fig. 108. Find the correct position and length of a short-circuited stub having the same characteristic impedance.

Solution (8).

$$V_r = V\frac{1-k}{1+k} = 0{\cdot}6V_f,\ \lambda = 3 \text{ metres}.$$

Set up V_f and V_r for current minimum as shown in Fig. 109 (lines AC, DCB). DCB is now rotated clockwise so that current leads, until the point B^1 is tangential. The vector diagram is then

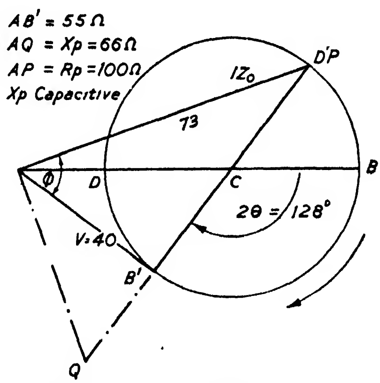


FIG. 109. Illustrating Problem 8.

completed, giving the condition required, namely that the vector $AP = IR_p = AD = I(Z_0)$. To arrive at this condition DCB has been moved through $2\theta = 128°$, the stub must be placed on the generator side of the current minimum a distance $\theta = 64° = {\cdot}18\lambda$, $= {\cdot}54$ metres, or $1{\cdot}54$ metre from the load.

From the vector figure we have:

$$\text{Impedance } Z = \frac{V}{I} Z_0 = 55\Omega.$$

From the impedance triangle

$$\begin{aligned} AB &= 40 \text{ units} &= Z &= 55\Omega \\ AP &= 73 \quad ,, &= R_p &= 100\Omega \\ AQ &= 48 \quad ,, &= X_p &= 66\Omega \end{aligned}$$

Since
$$X_p = Z_o \tan \phi$$

$$\tan \phi = \frac{66{\cdot}3}{100}, \ \phi = 33{\cdot}5^\circ.$$

$$\text{Length of stub} = \frac{33{\cdot}5 \times 3}{360} = {\cdot}28 \text{ metres.}$$

Problem (9). The conditions are the same as in the previous problem but it is desired to use an open-circuited stub having a characteristic impedance of 80Ω. Find its position and length.

Solution (9). Since an open-circuited stub has a capacitive reactance the position on the line must be towards the load from the current minimum, otherwise conditions are the same as in the last example. Thus the vector DCB will be rotated anti-clockwise by $2\theta = 128^\circ$ to bring the point B′ tangential to the reflection circle in the upper half-circle, so bringing the points D and P into the lower half as shown in Fig. 110. The stub is therefore $\theta = 64^\circ$ or

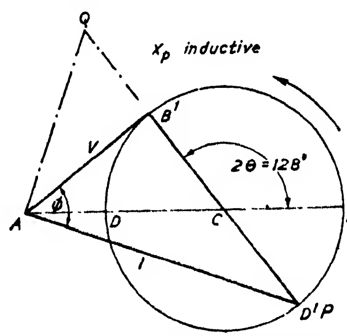


FIG. 110. Illustrating Problem 9.

${\cdot}18\lambda$ from the current minimum on the load side. X_p will be as before, viz. 66Ω, but inductive.

The reactance of the stub must be 66Ω, but since the characteristic impedance is now 80Ω, we have

$$X_p = Z_o \cot \phi, \quad \cot \phi = \frac{66{\cdot}3}{80} \quad \phi = 50{\cdot}4^\circ.$$

$$\text{Length of stub} = 0{\cdot}42 \text{ metres.}$$

It will be seen therefore that for a certain relationship between Z_1 and Z_0, or k, the distance between current minimum and the stub is the same whether a short-circuited or an open-circuited stub is used except in the former case the distance is

measured away from the load and in the latter case towards the load. Generally speaking short-circuited stubs are preferred as the end of the stub is then at low potential and can often be earthed.

The geometry of Figs. 109 or 110, i.e., when the point P is coincident with D', and the angle $AB'D'$ is therefore a right angle, enables us to find simple formulæ for the position and length of the stubs, when they have the same characteristic impedance.

The input impedance of stub $\frac{V}{I} \cdot Z_o = Z_o \frac{\sqrt{(V_f)^2 - (V_r)^2}}{2V_r}$

But $$V_f = V_r \frac{k+1}{k-1} \text{ and}$$

$\therefore$ Input impedance of stub $= \frac{Z_o}{2}\sqrt{\left(\frac{k+1}{k-1}\right)^2 - 1} = \frac{Z_o\sqrt{k}}{k-1}$

If stub is short-circuited type :

$$\frac{Z_o\sqrt{k}}{k-1} = Z_o \tan\phi \text{ or } \tan\phi = \frac{\sqrt{k}}{k-1} \quad . \quad . \quad (46)$$

If open-circuited :

$$\frac{Z_o\sqrt{k}}{k-1} = Z_o \cot\phi \text{ or } \cot\phi = \frac{\sqrt{k}}{k-1} \quad . \quad . \quad (47)$$

But from diagram $\frac{V_r}{V_f} = \cos(180° - 2\theta) = -\cos 2\theta.$

But $$\frac{V_r}{V_f} = \frac{k-1}{k+1} \text{ and}$$

Hence $$\cos 2\theta = \frac{1-k}{1+k} \quad . \quad . \quad . \quad . \quad (48)$$

where θ is measured away from the load for the shorted stub and towards the load for the open. These relationships are shown in Fig. 111, and enable stub position and length to be determined directly from the ratio I_{min}/I_{max}, provided the stub characteristic impedance is the same as the line.

The Impedance Circle Diagram

Suppose that we try to construct a diagram from which we can read off the impedance at any point on a line, if we know the value of k and the position of the current minimum.

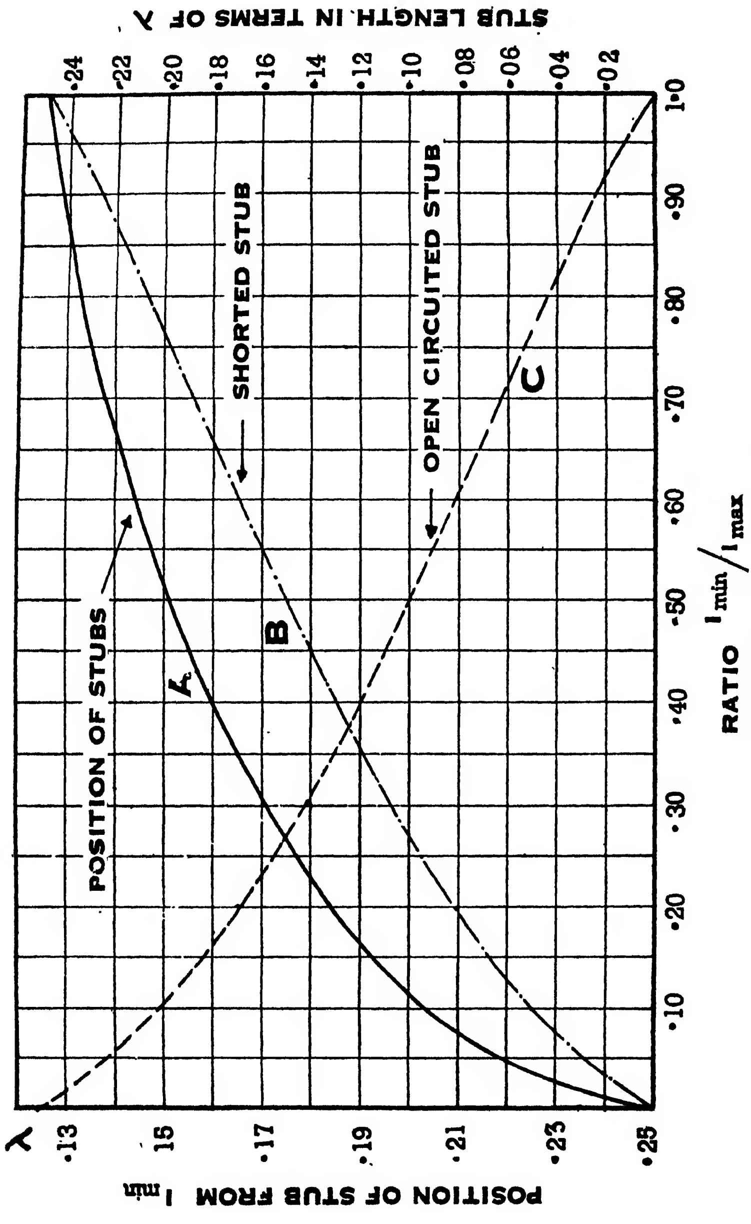


FIG. 111. Chart for Stub Matching.

It will be convenient to generalise by plotting all impedances divided by Z_0 and we will denote these "normalised" resistances and reactances by r and x respectively. All lengths along the line will be measured by the electrical or angular length, θ, from the current minimum.

We know that at the current minimum the voltage is $V_f + V_r$ whilst the current is $\dfrac{V_f - V_r}{Z_0}$ and that the impedance is a pure resistance given by $\dfrac{V_f + V_r}{V_f - V_r} Z_0$.

Hence
$$r_{max} = \frac{V_f + V_r}{V_f - V_r} = \frac{1}{k} \qquad . \quad . \quad . \quad (49)$$

Similarly, at a current maximum, $r_{min} = k$ . . . (50)

If k is 0·5, for example, then at the current minimum the vector for the impedance is of length 2 measured along the resistance axis, whilst at the current maximum it is 0·5, in the same direction.

It can be shown that the locus of the vector for the impedance at all points along the line is a circle, with its centre on the r axis and having a diameter $r_{max} - r_{min}$.

We have now to see how P moves round this circle as θ varies. Movement round half the circle clearly corresponds to movement along the line of 90° ($\lambda/2$) but the circle is not evenly divided.

Some of the conditions can readily be established. Suppose we consider an open-circuited line, then the current minimum is at the end of the line and the impedance at any point measured back from the end is known to be $-jZ_0 \cot \theta$. Hence $x = -j \cot \theta$ and we can therefore prepare a scale of θ along the x axis. For example, if $\theta = 45°$, $x = -1$.

Now if the line is correctly terminated, the impedance at all points will equal Z_0 (or $r = 1$). Since the diagram must cover this case, all the lines for constant distances from the current minimum must pass through 1, as well as intersecting the x axis, as already discussed.

It can be shown that these curves are also circles, of which three specimens are drawn on Fig. 112. It will be seen that the 90° curve is formed by the r axis. A consideration of the

open-circuited line (current minimum at end) will show that we follow clockwise round the circle when going from a current minimum *towards* the generator. If we have a different value

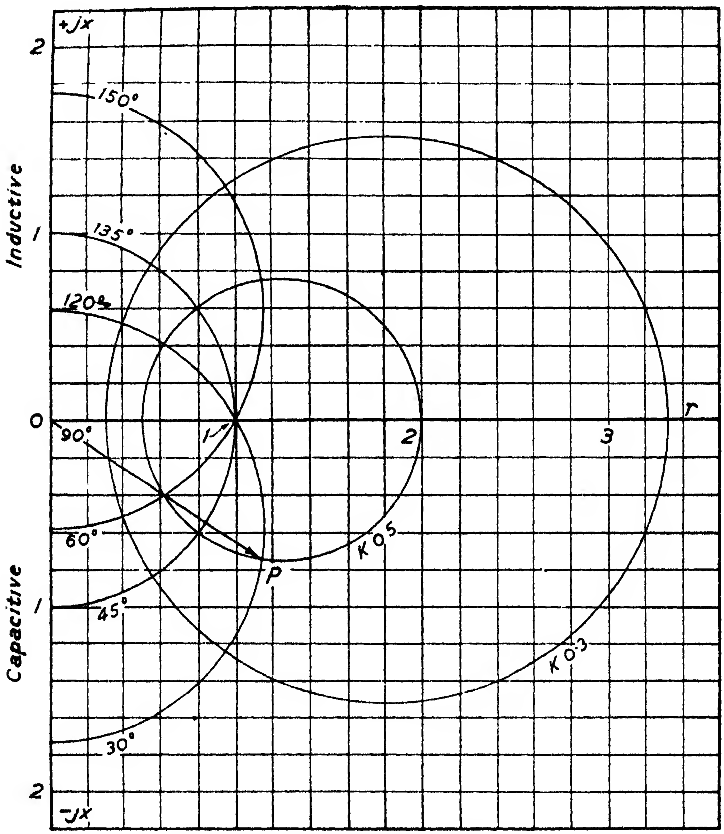


Fig. 112. Impedance Circle Diagram.

for k, then we follow another circle, such as that shown for $k = 0{\cdot}3$.

To find the impedance of the line at a certain distance from the current minimum (Z_o and k being known) we shall pass round the appropriate k circle until we intersect the required θ circle and then read off the values of r and x. Conversely, if we wish to know what value of k will result from the use of a certain load, we can (after " normalising ") find upon which

circle the points lie. The diagram can also be used to find the correct length and characteristic impedance for a matching line.

The Admittance Circle Diagram

A similar diagram can be prepared giving susceptance b and conductance g instead of x and r and this is convenient for stub-matching problems, where we are dealing with stub and line in parallel.

At a current maximum, the admittance, y, is a pure conductance, the "normalised" value, g_{max}, being $\frac{1}{k}$ and at a

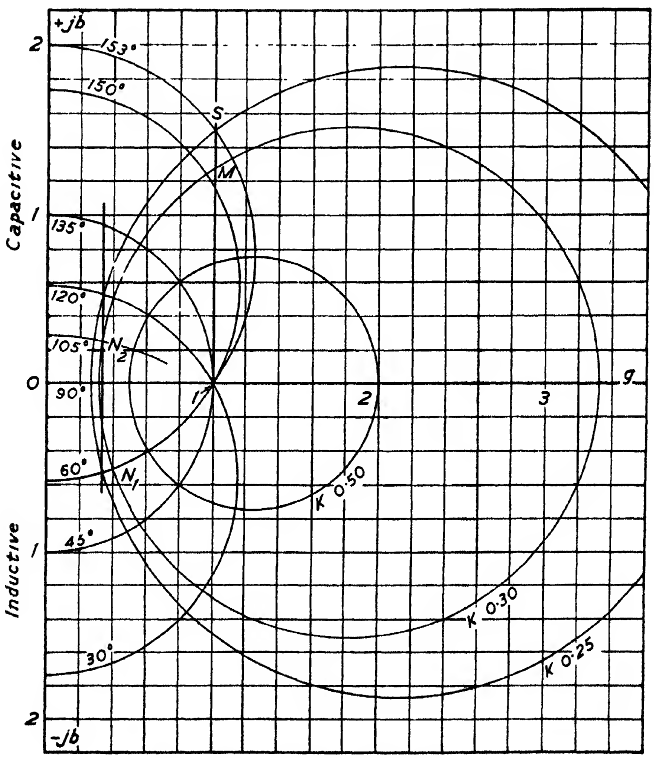


FIG. 113. Admittance Circle Diagram.

current minimum, $g_{min} = k$. In this case, again, it can be shown that y lies on a circle, as shown in Fig. 113. The circles for constant distances are the same shape as before but are now measured from a current maximum.

The circle for $k = 0{\cdot}25$ has been drawn, in order that Problem (8) can be checked. Since the stub can only affect the susceptance of the line it must be placed at a point where the normalised conductance is unity and must have a susceptance equal and opposite to that of the line at this point. The susceptance of the short-circuited stub (if less than $\lambda/4$ long) will be negative, and hence the position of the stub is 153° from the current maximum and its susceptance is 1·5℧. Since the normalised reactance of the stub is j tan θ_s its normalised susceptance will be $-j$ cot θ_s and hence $\theta_s = 33{\cdot}7°$ and its length 0·28 m. These results agree with those previously found.

Double-stub Matching

It is often not convenient, especially with concentric lines, to have to move the position of the stub, although the length of the stub in use can be conveniently changed. If we use two stubs

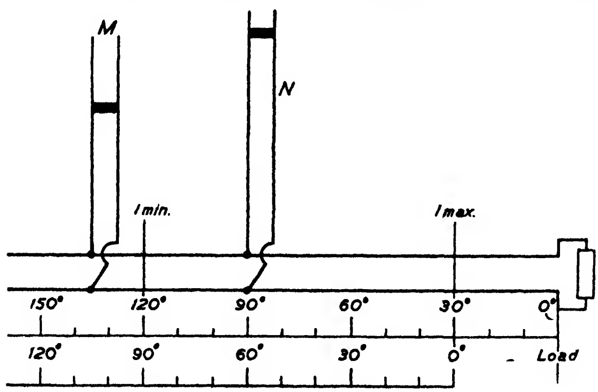


FIG. 114. Double Stub Matching.

at a fixed distance from the load and with a fixed spacing between them, as shown in Fig. 114, then a wide range of loads can be matched to the line by adjusting the lengths of the two stubs. The stubs must not be $\lambda/2$ apart because this spacing merely puts them in parallel, in effect. A common spacing is $\lambda/8$.

The necessary conditions can be examined on the admittance circle diagram and will be best understood by reference to a specific case.

Problem (10). A concentric line with air dielectric, has $Z_0 = 80\Omega$ and is being used at 1,000 Mc/s. When the load is connected it is found that $k = 0{\cdot}25$ and that the first current minimum occurs at 10 cm from the load.

Matching is to be carried out by two short-circuited stubs, one 7·5 cm from the load and the other 11·3 cm. Find suitable lengths for the two stubs, which have $Z_0 = 80\Omega$.

Solution (10). $\lambda = 30$ cm. Hence, in terms of angles, distances are as shown in Fig. 114. Thus the current minimum is 120°, the first stub 90°, and the second approximately 136° from the load. Referring now to the admittance diagram of Fig. 113 and working back from the current maximum, we follow round the $k = 0{\cdot}25$ circle to the intersection with the 60° circle.

Here the stub N adds susceptance but the conductance remains the same and hence we move up the vertical line N_1, N_2. N_2 has got to be chosen so that, if we move round the k circle passing through it, through a line length of 45° (the stub spacing) we reach the vertical line through 1. Then stub M can put in the necessary susceptance to make the admittance just to the left of M equal to G_0, so that the main line is correctly terminated.

If we have a complete mesh of k and distance circles, this can be done by inspection. In Fig. 113 it will be seen that the $k = 0{\cdot}3$ circle is approximately correct.

Just to the right of N the susceptance is seen to be $-\ 0{\cdot}52\mho$, whilst just to the left it is $+\ 0{\cdot}25\mho$. The susceptance of the stub is, therefore, $+\ 0{\cdot}77\mho$. If θ_N is the angular length of this stub, then $-j\cot\theta_N = j\,0{\cdot}77$ or $\theta_N = 128°$ and the actual length of N is 10·7 cm.

Similarly, the susceptance just to the right of M is $+\ 1{\cdot}25\mho$ and M must provide $-\ 1{\cdot}25\mho$. Hence $\theta = 38{\cdot}6°$ and the actual length of M is 3·2 cm.

A Note on Lines with Losses

The theory of lines when conductor resistance R and leakage conductance G between conductors are allowed for is given in a number of textbooks and the results of most importance in connection with radio-frequency lines will be set forth here, without proof.

The characteristic impedance Z_0 becomes $\sqrt{\dfrac{R + j\omega L}{G + j\omega C}}$ whilst the propagation constant (the equivalent of m in our equations) is

$$P = \sqrt{(R + j\omega L)\,(G + j\omega C)}.$$

It will be seen that P is complex quantity and may be written $\alpha + j\beta$ where α is called the attenuation constant, because the maximum (or R.M.S.) value of the voltage at a point x along a correctly-terminated line is related to that at the generator by the equation

$$E_x = E_g e^{-\alpha x} \quad . \quad . \quad . \quad . \quad (51)$$

Similarly

$$I_x = I_g e^{-\alpha x} \quad . \quad . \quad . \quad . \quad (52)$$

It will be seen that the voltage decreases exponentially along the line, instead of being constant in value, as in the case of a "loss-free" line.

In the case of lines used at radio frequencies, $Z_o = \sqrt{\frac{L}{C}}$ is a very good approximation and

$$\alpha = \frac{R}{2Z_o} + \frac{GZ_o}{2} \quad . \quad . \quad . \quad . \quad (53)$$

The loss in a correctly-terminated line is therefore given by

$$\text{Loss in decibels} = 20 \log_{10} \frac{I_g}{I_r} \text{ or } 20 \log_{10} \frac{E_g}{E_r}$$

$$= 20 \log_{10} e^{\alpha l} = 8{\cdot}686 \alpha l \quad . \quad . \quad . \quad (54)$$

When R and G are allowed for, the equations giving the voltage and current at any point along lines not correctly terminated may be written in terms of hyperbolic functions instead of the circular functions in the "loss-free" line equations.

In the case of a line $\lambda/4$ long and open-circuited at the far end the impedance placed across the generator is a resistance given by

$$Z_{oc} = Z_o \tanh \alpha l \quad . \quad . \quad . \quad . \quad (55)$$

whilst in the case of a $\lambda/4$ line, short-circuited at the far end,

$$Z_{sc} = Z_o \coth \alpha l \quad . \quad . \quad . \quad . \quad (56)$$

The imaginary part, β, of the propagation constant determines the phase of the voltage or current along the line, so that the complete expression for the voltage at x is

$$E_x = E_g e^{-\alpha x} (\cos \beta x - j \sin \beta x) \quad . \quad . \quad (57)$$

which is seen to be a vector of length $E_g e^{-\alpha x}$ and having a phase angle βx with respect to the generator voltage.

Since values of E_x will repeat when βx is increased by 2π, it follows that the wavelength λ is $\frac{2\pi}{\beta}$ and hence the velocity of propagation is ω/β.

The velocity now depends upon frequency but the difference is usually small over the range of radio frequencies. The velocity is always less than that of a wave in " free space " even when the dielectric is almost entirely air. The percentage reduction of velocity is sometimes referred to as " velocity slip."

Calculation of Losses in Concentric Cables

The conductor losses may be calculated with fair accuracy, using a formula developed by Russell for the A.C. resistance of concentric tubes. For radio frequencies this becomes

$$R_f = \sqrt{\mu f \rho}\left(\frac{1}{r_1} + \frac{1}{r_2}\right) \text{ E.M. units per cm.}$$

If the conductor is copper (resistivity 1·7 microhms per cm. cube) and the dielectric has unit permeability, then

$$\text{Resistance per km} = 41{\cdot}2 \times 10^{-4}\sqrt{f}\left(\frac{1}{r_1} + \frac{1}{r_2}\right) \text{ ohms.} \qquad (58)$$

If conductor losses only are taken into consideration, $\alpha = \frac{R}{2Z_0}$

or, in terms of line dimensions,

$$\frac{41{\cdot}2 \times 10^{-4}\sqrt{f}\left(\frac{1}{r_1} + \frac{1}{r_2}\right)}{2 \times 138 \log_{10}\frac{r_2}{r_1}} \qquad (59)$$

from which, attenuation in decibels per km.

$$= 1{\cdot}30 \times 10^{-4}\sqrt{f}\,\frac{\left(\frac{1}{r_1} + \frac{1}{r_2}\right)}{\log_{10}\frac{r_2}{r_1}} \quad . \quad . \quad (60)$$

Let us now assume r_2 to be fixed and investigate the effect of varying r_1 denoting the ratio by x.

Then equation (60) may be written (for a fixed frequency).

$$\text{Attenuation in decibels per km} = y = \frac{k(x+1)}{\log_{10} x}$$

and this will be a minimum when $\frac{dy}{dx} = 0$

that is, when $\log_{10} x = \frac{x+1}{x}$

This equation is solved by the value 3·6.

In the curve of Fig. 115, $\frac{x+1}{x}$ is plotted against x so that the variation of attenuation with ratio can be seen. The curve

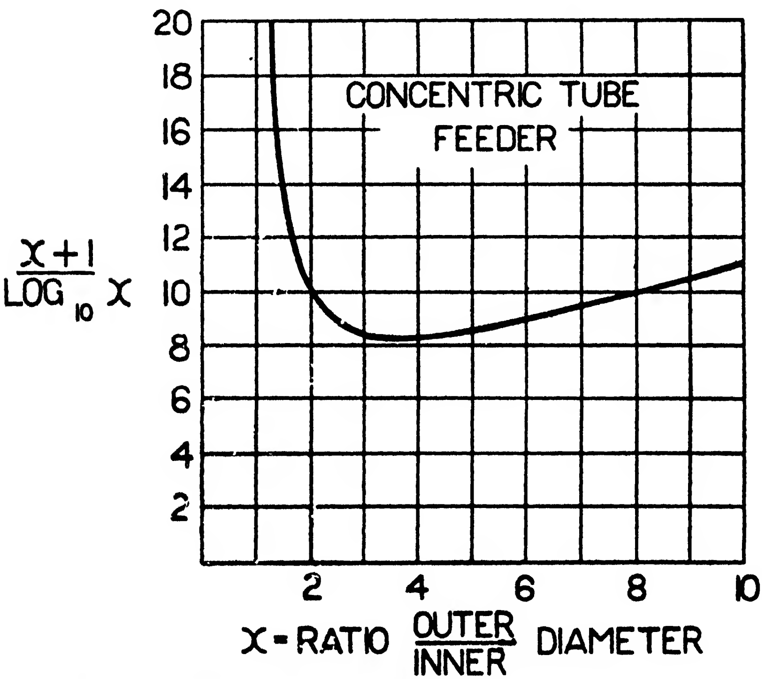


FIG. 115. Variation of Loss with Ratio of Conductor Diameters.

shows the great increase in attenuation if too small a ratio is used.

The resistance of parallel-wire lines may be worked out by considering each wire separately, using equation 58 and putting $r_2 = \infty$. This method ignores earth current losses

and the proximity effect of the other wire, this latter effect being usually small. By a similar procedure to that adopted above for concentric lines, in the case of twin parallel lines,

$$\text{Attenuation} = \frac{1{\cdot}30 \times 10^{-4}\sqrt{f}}{r \log_{10}\frac{d}{r}} \text{ db/km} \qquad (61)$$

It is important to realise that if the characteristic impedance of a line is low, then it must also have a low resistance if power is to be transmitted along it with good efficiency.

Suppose a power P is to be delivered by a line of characteristic impedance Z_0 and conductor resistance r. Then the line current is $\sqrt{P/Z_0}$ and voltage $\sqrt{PZ_0}$. Conductor losses are $P\frac{r}{Z_0}$ and hence, if these only are considered, equal transmission efficiencies will be obtained from different types of line having the same value of $\frac{r}{Z_0}$.

The line having the larger Z_0 will tend to have larger dielectric losses, because the voltage will be higher.

Types of Radio-frequency Concentric Lines

The concentric lines employed at radio stations fall into three main classes:

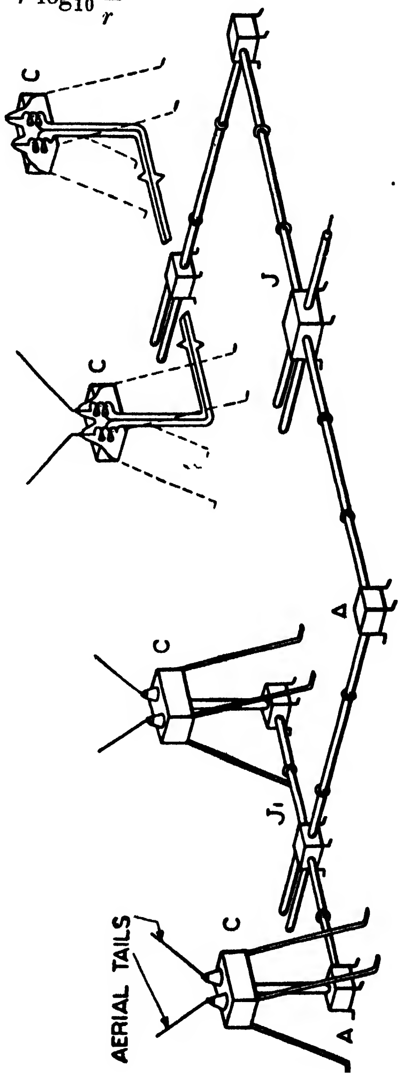


FIG. 116. Early Type of Concentric Line Lay-out.

(a) Lines having rigid tubular conductors.
(b) Flexible cables having spaced insulators.
(c) Flexible cables having continuous solid insulation.

Type (a) was the first to be developed, being used both at the transmitting and receiving stations in the original beam system. It is still widely used at transmitting stations. The outer tube is usually supported at frequent intervals a short distance above ground on iron stakes, and special boxes are used to negotiate bends and junctions and to provide for expansion. The general layout of a Marconi feeder system for a broadside array is shown in Fig. 116. At junctions it will be necessary to have an impedance-matching arrangement, because the two branches in parallel will have an impedance of only $\frac{Z_o}{2}$. The reactance transformer described on page 184 is usually used.

In early types of line, expansion of the outer tube was

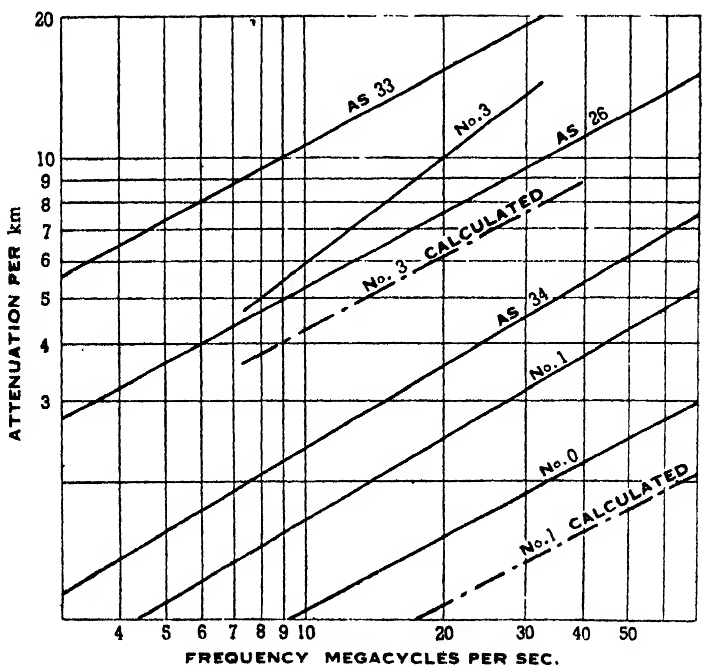


FIG. 117. Losses in Concentric Lines.

allowed for by the insertion of corrugated diaphragm connectors at intervals, but owing to the electrical discontinuity this creates, the expansion of the outer tube is now allowed for by sliding sleeves in the same way as that of the inner tube.

Particulars of some typical rigid, concentric lines are given in Table IX, whilst Fig. 117 shows measured and calculated attenuations.

TABLE IX. *Particulars of some Rigid Concentric Lines*

Inner and outer conductors both of copper.

Attenuations and efficiencies are calculated values at 20 Mc/s.

Type.	Outer Radius of Inner Tube Cm.	Inner Radius of Outer Tube Cm.	Ratio.	Percentage Efficiency per Km.	Attenuation Decibels per Km.	Z_0
No. 0	3·30	13·0	4	81·3	0·95	83·0
No. 1	1·11	4·44	4	77·8	1·09	83·0
No. 2	0·875	3·17	3·6	70·5	1·52	77·0
No. 3	0·238	0·795	3·34	24·8	6·05	77·2

It will be seen from Fig. 117 that with large concentric-tube lines the total loss is about double the copper loss, whereas with small lines the total loss is less than double because the copper loss is greater for the smaller conductor.

Concentric-tube feeders are sometimes designed to be buried below the ground to reduce expansion troubles and give more perfect screening. In certain cases arrangements are made to pump in dry air or an inert gas under pressure, in order to exclude moisture and maintain the characteristics of the line constant under varying atmospheric conditions. Whether such a refinement is necessary is doubtful, as the experience of Messrs. Cable and Wireless over a large number of years with ordinary concentric-tube feeders above ground is that they give no trouble and that their maintainence costs are practically nil, except for the painting of the supporting irons.

It is important to realise that in a concentric line the voltage gradient, that is, the voltage across a centimetre of the dielectric, is greatest at the surface of the conductor. This is because the capacitance of a thin concentric ring of the dielectric is least at the surface, and, since the same charge is flowing through all

these concentric rings in series, the voltage across them is inversely proportional to their capacitance.

It can be shown that the potential gradient at the surface of the inner conductor is given by

$$g = \frac{V}{r_1 \log_e \frac{r_2}{r_1}} \text{ volts/cm} \quad . \quad . \quad . \quad (62)$$

The potential gradient must not be too high or a discharge will take place from the surface of the conductor. At power frequencies the potential gradient at which this takes place is about 30 kV/cm but at radio frequencies it is somewhat lower and corona is much more violent.

When annular insulators are used to support the inner conductor it is very important that they should fit tightly to the inner conductor because the potential gradient across any air-gap at the surface would be very high, owing to the fact that the insulator has a higher dielectric constant.

In the last few years a variety of flexible and semi-flexible feeder cables have been developed, the former for receiver work and the latter of size large enough for transmission, or for reception where very low loss indeed is required. The advantages of the larger size over rigid types are that they can be handled like ordinary electric-power cables and the jointing involves much the same technique, whilst the overall characteristics tend to be more free of discontinuities.

One type consists of a solid copper wire as central conductor spaced concentrically within a lead outer tube by means of spacing discs made from a synthetic resin, known as Trolitul, placed about 3 inches apart. Trolitul has a high dielectric constant and a low power-factor and has the advantage of not being hygroscopic. The spacers are assembled on the inner copper wire and the lead extruded over the discs from a lead press, and although there is no practical limit to the length that can be built in one operation, it is usual to manufacture such cable in lengths up to 500 metres. After sheathing, the cable is armoured in a normal way with steel tape and jute covering and even the largest size can be rolled on to drums for transport and installation direct into trenches. The installation of the cables follows power-cable technique. Thus for

jointing the centre wire will be scarfed, soldered, bound with fine wire and resoldered, and the outer lead will be joined with a lead sleeve with wiped joints, and where the joint is exposed the join will be enclosed in a compound-filled, cast-iron joint-box, as shown in Fig. 118. For the termination of such a cable a special end-fixing gland has been designed as shown in Fig. 119.

Although for a given diameter of outer, a lead outer is not quite so efficient as a copper, ease of manufacture and installation are strong recommendations for its use on low powers.

Cables with solid dielectric are now in common use at high radio frequencies. This is mainly due to the development of

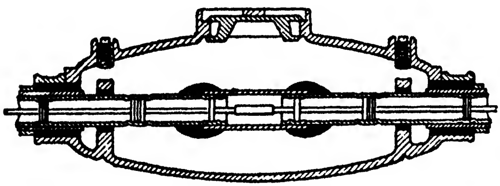


FIG. 118. R.F. Cable Joint Box.

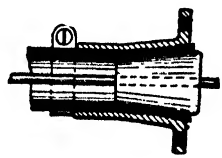


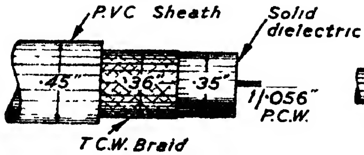
FIG. 119. R.F. Cable Termination.

the thermo-plastic polythene, having a value such as 0·0003 for tan δ (δ is the loss-angle, that is, the complement of the phase angle), whilst its dielectric constant is about 2·3. It is mechanically tough and flexible, even at temperatures below zero.

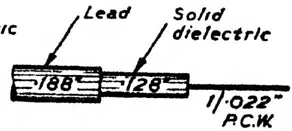
Assuming that tan δ remains constant with frequency, the loss in a dielectric will be proportional to frequency, since there is a constant loss of energy per cycle. Conductor loss, on the other hand, has been shown to be proportional to $\sqrt{f}$. Hence, if we wish to use a cable at higher frequencies, the importance of having a very low value for tan δ will increase. For example, even with a polythene-insulated cable, the dielectric loss at 10,000 Mc/s would be 57% of the total loss, although at 1 Mc/s it would only be 1%.

Twin types of cable are also available, copper-wire braiding often being included as a screen for both conductors.

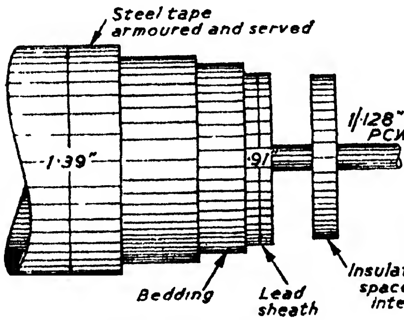
Table X gives particulars of a few types of cable, the



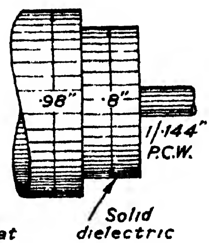
Uni-Radio, No. 1.



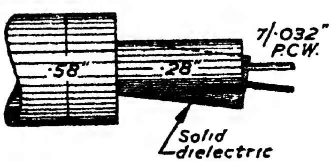
Uni-Radio, No. 33.



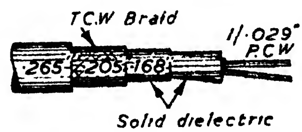
Uni-Radio, No. 8.



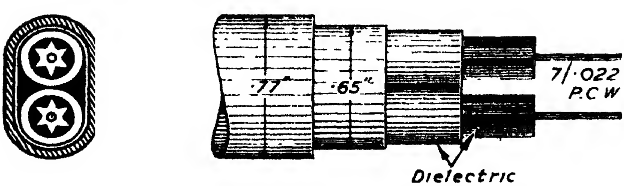
Uni-Radio, No. 10.



Du-Radio, No. 29.



Du-Radio, No. 28.



Du-Radio, No. 20.

TCW Tinned Copper Wire.
PCW Plain Copper Wire.
PVC Polyvinol Chloride.

Fig. 120. Types of Radio-Frequency Cables.

* TABLE X. *Some Typical Flexible R.F. Cables*

Service Type.	Uni-radio (Concentric).				Du-radio (Twin).		
	No. 1.	No 8	No. 10.	No. 33.	No. 20	No. 28	No. 29.
Mean Z_0 (ohms)	71	103	69	72	150	98	125
Att. db./100 ft.:							
10 Mc/s.	0·49	0·17	0·24	1·4	1·0	1·9	0·36
100 ,,	1·74	0·58	0·88	4·6	3·3	6·1	1·1
1,000 ,,	8·0	2·0†	3·9	16	—	22	6·5
Mean power rating (watts):							
10 Mc/s	2,200	4,500	7.600	500	—	470	5,100
100 ,,	640	1,440	2,060	150	—	143	1,445
1,000 ,,	140	350	470	46	—	42	350
Velocity ratio	0·67	0·96	0·67	0·67	—	0·67	0·67

* Extracted by permission of the Inter-Services R.F. Cables Committee.

† Cable of this type cannot normally be used at this frequency, because of reflections at the insulators.

construction of which is shown in Fig. 120. These particulars are given by permission of the Inter-Services R.F. Cables Committee and refer to Service types of cable, but similar types are available commercially from several firms.

Parallel-wire Radio-frequency Lines

The construction of an open-wire line for radio-frequency work is similar to that used for telephone lines, except that the lines are usually hung from a cross-arm on long, rod insulators having small dielectric loss.

The spacing between wires depends upon a number of considerations. There is the obvious practical requirement that there must be no risk of the conductors touching when swaying in a high wind. In fact, for transmission work they must not come within a distance apart which will be small enough to cause corona. For low-power transmission a spacing of about 15 cm is usual, whilst for higher powers the spacing may go up to 30 cm, above which the cross-arms, etc., become rather clumsy.

The larger the spacing, the more likelihood there is of the line radiating or picking up radiation. In the case of lines connected to receivers this is particularly important and a spacing of about 5 cm. is frequently used. With this small spacing it is necessary to introduce separators every 3 m, or less.

Increasing the gauge of wire employed has little effect upon Z_0 (as can be seen from Table XI), but the potential gradient at the surface of the conductor varies approximately as the inverse of the diameter. Also, of course, we are increasing the periphery of the conductor which will carry the high-frequency currents. Hence, for higher powers, heavier conductors will be used. Generally speaking, No. 6 gauge wires (radius 0·243 cm), separated 30 cm, represent the upper practical limit for two-wire lines and this size will be satisfactory for a transmitter producing a 50 kW carrier and capable of 100% modulation.

For powers in excess of 50 kW, four-wire lines may be used. The characteristic impedance is lower and hence the voltage for a given power is reduced. Four-wire lines may be constructed with the go-and-return pairs disposed at opposite

TABLE XI. *Particulars of some Two-wire Lines*

Wire Gauge No.	Radius of Wire (Cms.)	Distance between Wires (Cms.)	Ratio d/r.	Percentage Efficiency per Km.	Attenuation Db/Km at 20 Mc/s.	Z_0 ohms.
		10	41·1	71·1	1·48	445
6	0·243	20	82·2	75·0	1·25	530
		30	123	76·9	1·14	578
	—	10	49·3	67·8	1·69	468
8	0·203	20	98·6	72·8	1·43	551
	—	30	148	73·8	1·32	600
	—	10	75·7	58·3	2·34	519
12	0·132	20	151	63·0	2·01	602
		30	227	65·2	1·86	650

sides of a rectangle, or diagonally (see Fig. 89). The former gives a lower Z_0 but does not make such an easy mechanical layout. Because of the layout, there are no discontinuities and it is found that such lines are more easily lined-up than the two-wire. A four-wire line using No. 6 gauge wires and 30 cm spacing could be used with a telephone transmitter of 100 kW carrier power.

Comparison between Concentric and Parallel-wire Types of R.F. Line

There is a divergence of opinion regarding the relative merits of concentric and parallel-wire lines. Both can be made about equally efficient and the comparison is mainly on the grounds of cost and convenience.

The parallel-wire line is suitable for direct connection to aerials which are balanced with respect to earth, such as the horizontal types (see page 281), whilst the concentric line is evidently suited for connection to aerials which are earthed.

The concentric type has the great advantage that it does not radiate. There is necessarily radiation from the parallel-wire type and this has to be kept down as much as possible by adjusting the line and its termination very carefully so that stationary waves are reduced to a minimum and hence the current is the smallest possible. Care must also be taken to balance the load with respect to earth, so that the current in each wire is the same, as evidently any excess current in one

wire not balanced by a return current in the other will cause radiation to take place from the line.

In the case of lines for high-power transmitters, the concentric type requires considerably more material and costs more to install but is less critical to adjust, and its characteristics are more constant with weather conditions. It can be installed just above or under the ground out of the way of aerials. The parallel-wire type is more suitable where the installation is at all temporary and in the event of a flash-over the trouble is more quickly located and repaired.

For low-power transmitting, or for reception, the flexible concentric types are very convenient. The fact that the concentric type does not pick up stray E.M.F.'s, even if not correctly adjusted, is of very great importance in reception.

Troubles due to radiation naturally increase with frequency and therefore concentric lines are almost exclusively used at the very high frequencies.

Effect of Insulator Spacing

All our calculations assume a uniformly distributed capacity between the lines but solid insulation must be employed to

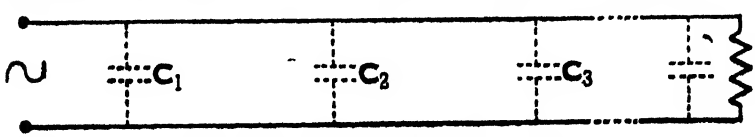


FIG. 121. Effect of Insulator Spacing.

keep the conductors in position and these insulators will, in most cases, be spaced at intervals in order to reduce dielectric loss and cost. Since all known materials have a dielectric constant greater than that of air, we are therefore inserting additional parallel capacities at intervals, which may have values up to about $5\mu\mu$F.

The effect of the insulators is very dependent upon their spacing in relation to the wavelength being used.

Consider a feeder (Fig. 121) in which the insulators are spaced $\frac{\lambda}{2}$ apart. It can be seen from (34) that if $\theta_1 = \pi$, then $Z_s = Z_r$. That is, a $\frac{\lambda}{2}$ line does not transform impedances,

but a terminating Z_r produces an impedance Z_r across the generator. Hence C_3 is effectively in parallel with C_2 to give $2C$ and this will produce the equivalent of $3C$ at C_1, so that if there are n insulators along the feeder, a capacity of nC is placed across the generator, and the input impedance is very different from Z_0 even if the feeder is correctly terminated; and will depend greatly upon the length of the feeder. There will be large reflected waves set up at the nearer insulators where low capacity reactances are effectively shunted across the feeder. Fig. 122 shows the variation of input impedance

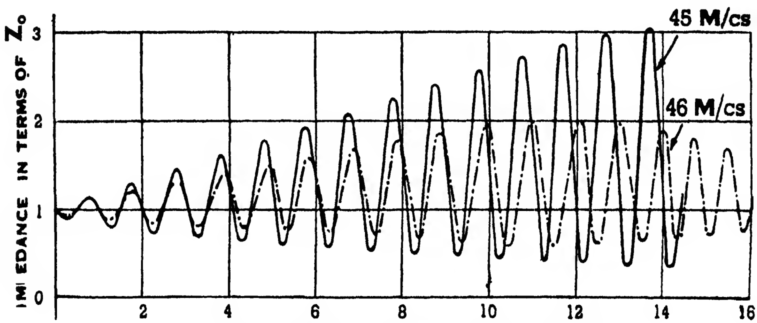


FIG. 122. Effect of Insulators on Input Impedance.

looking towards the terminal load as the distance along the feeder is increased. If insulators are spaced at $\frac{\lambda}{4}$ intervals, the capacity reactance of C_1 becomes an inductive reactive at C_2, partially cancelling the effect of C_2, and so on. No cumulative effects occur and the variation of input impedance with length is therefore small. Such an arrangement would be, however, critical with frequency. It is more usual, therefore to increase the number of insulators considerably, whereby the change of input impedance with length can be reduced to a negligible amount, eleven per wavelength being a usual figure. This question of insulator spacing principally concerns ultra-short wave feeders because on the longer wavelengths such spacings as $\frac{\lambda}{2}$ or $\frac{\lambda}{4}$ would be quite inadequate mechanically.

Matching over a Wide Frequency-band

All the methods so far discussed are designed to work at a specified frequency. If the modulation frequency concerned is a comparatively high fraction of the carrier frequency, then it is clearly desirable that the match should be as good as possible for the whole of the side-band frequencies. There will also be cases where it is inconvenient, or impossible, to adjust the line matching when changing the frequency of transmission.

The amount by which the matching of quarter-wave lines, stubs, etc., is disturbed by a change of frequency is mainly affected by the degree of mis-match which is being corrected for. If, for example, a quarter-wave line is being used between

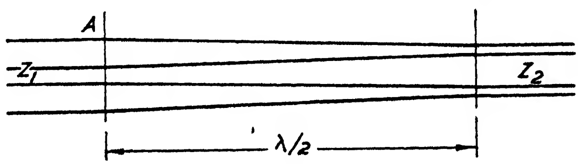


FIG. 123. Concentric Line with Linear Taper.

a load which is very different from the Z_0 of the main line, then a small change of frequency will produce serious reflection.

If it is desired to retain the match over a band of frequencies, it will be better to do the matching in two steps—that is, to have two $\frac{\lambda}{4}$ lines in series. The more $\frac{\lambda}{4}$ lines used in series, the wider would be the bandwidth over which the matching would remain approximately correct.

This leads us to the consideration of a tapered line, in which Z_0 is changing continuously along its length. By this means it is possible to match approximately over a very wide band. An example of a linearly-tapered concentric line is shown in Fig. 123. The matching section should be $\frac{\lambda}{2}$ long, where λ is the wavelength at the middle of the band to be covered. As an example of what can be achieved, in a certain case Z_2 was 70Ω and Z_1 was 120Ω. For λ the impedance at A was 120Ω, so that there were no standing waves, whilst for $0{\cdot}6\ \lambda$, k was only $0{\cdot}83$.

Exponential taper is also used, that is, a line in which Z_0 varies exponentially with length of matching line. A matching line approximating roughly to an exponential line is shown in Fig. 124 which was used for connecting a 4-wire line having a Z_0 of 300Ω to a 2-wire line for which Z_0 was 600Ω. For this, k was 0·82 at 20 Mc/s and 0·83 at 32 Mc/s.

Stubs can also be used to improve the performance of a line and aerial system over a frequency band. We shall see later (page 260) that if the frequency applied to a $\frac{\lambda}{2}$ aerial rises above the resonance value, the aerial has inductive reactance. Suppose that at a point approximately $\frac{\lambda}{2}$ back from the aerial, we place a short-circuited $\frac{\lambda}{4}$ stub. Then at the resonance frequency this will put an infinite impedance across the line and will, therefore,

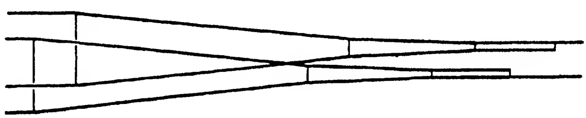


FIG. 124. Four-Wire to Two-Wire Matching by Exponential Taper.

have no effect. If a higher frequency is applied, however, it will put a capacitance across the line. Since the inductance introduced by the aerial has been transferred nearly unchanged to the stub junction, this inductance is partially neutralised by the stub capacitance.

If we wish to make a junction in a line system, we can ensure matching for all frequencies by making the line up to the junction of twice the Z_0 of the two branch lines. For example, the main line might be a 4-wire line having a Z_0 of 300Ω and the branches 2-wire lines of 600Ω.

The best possible matching up of aerial and line systems is necessary when long lines are used in connection with television, pulse modulated or radar transmitters. The time taken for a wave to travel down the line may be comparable with the time taken to transmit an element of the picture, or the duration of a pulse. If there is reflection at the aerial end and again at the transmitter end, a wave may reach the aerial after the

main transmission and result in a second transmission, or echo.

Thus at a television station it may be necessary to balance out all irregularities in the line to the aerial caused by junction boxes, etc., and arrange the insulator spacing very carefully.

The Q of a Resonant Line

When dealing with a resonant circuit consisting of a coil and a condenser, Q is usually defined as the ratio of the voltage across the condenser at resonance, to the E.M.F. injected into the circuit, and is given by $\omega . L/R$. A more general definition which can be applied to any circuit capable of resonance, such as short lengths of line, wave guides or cavity resonators, will now be given.

$$Q = \frac{\omega \times \text{Energy stored per cycle}}{\text{Energy dissipated per second}} \quad . \quad . \quad (63)$$

If we apply this definition to the ordinary LC circuit, then when the current is a maximum the energy stored in the form of a magnetic field is given by $\frac{1}{2} LI^2_{max} = LI^2_{R.M.S.}$. The energy dissipated per second is $I^2_{R.M.S.}R$ and therefore

$$Q = \frac{\omega_o LI^2_{R.M.S}}{I^2_{R.M.S.}R} = \frac{\omega_o L}{R}, \text{ as before.}$$

The Q of a circuit is also given by reactive volt-amperes/watts, since the reactive volt-amperes for the coil are $\omega_o LI(I)$ and the watts I^2R, again giving $Q = \omega_o L/R$.

Consider a line having a capacitance of C farads per metre, a characteristic impedance of Z_o ohms and an attenuation of α per metre. The line is short-circuited at one end and a p.d. of E_g (R.M.S.) at a frequency f, which makes the line exactly $\frac{\lambda}{4}$-long, is applied at the other.

The error introduced by considering the voltage and current distribution to be the same as in the loss-free line will be very small. The voltage at a point given by θ from the generator will be $E_g \cos \theta$. Considering an element of the line of electrical length $d\theta$, its capacitance will be $Cd\theta \frac{\lambda}{2\pi}$ and, therefore, at the instant when the voltage is a maximum, the energy stored in the electrostatic field will be

$$\frac{1}{2}\left(C \,.\, d\theta \,.\, \frac{\lambda}{2\pi}\right)(\sqrt{2}E_g \cos\theta)^2 = E_g^2\frac{C\lambda}{4\pi}(1 + \cos 2\theta)\, d\theta \text{ joules.}$$

The current is zero at all points along the line at this instant and therefore there is no stored magnetic energy and the total stored energy is

$$E_g^2\frac{C\lambda}{4\pi}\int_0^{\frac{\pi}{2}}(1 + \cos 2\theta)d\theta = \frac{E_g^2 C\lambda}{8} \text{ joules.}$$

The input impedance of a $\frac{\lambda}{4}$ line, short-circuited at the far end, is a pure resistance $Z_{sc} = Z_0/\tanh \alpha l$, and, since αl is small for R.F. lines $\frac{\lambda}{4}$ long, this may be approximated to $\frac{Z_0}{\alpha l}$.

The power input is $\therefore \quad \dfrac{E_g^2 \alpha l}{Z_0}$

and $$Q = \frac{\omega C\lambda Z_0}{8\alpha l} = \frac{\omega C Z_0}{2\alpha}, \text{ since } l = \frac{\lambda}{4}.$$

But $$Z_0 = \sqrt{\frac{L}{C}}, \text{ hence } Q = \frac{\pi f\sqrt{LC}}{\alpha} \quad . \quad . \quad (64)$$

and, since $\frac{1}{\sqrt{LC}} = v$, the velocity of propagation,

$$Q = \frac{\pi f}{\alpha v} \quad . \quad . \quad . \quad . \quad (65)$$

When the dielectric is air, v can be put approximately equal to $c = 3 \times 10^8$ m. per second.

As an example consider a length of No. 1 concentric line, as specified in Table IX, used as a $\frac{\lambda}{4}$ resonator at 60 Mc/s. The attenuation at 60 Mc/s is approximately 2 db per km (see Fig. 117) from which α is $2{\cdot}31 \times 10^{-4}$ per m. Hence the Q is 2,720. The input resistance at resonance will be 359,000Ω so that the line will act as a parallel-resonant circuit of very high "dynamic" resistance.

A similar argument for an open-circuited line of similar constants shows that Q will be the same whichever way it is used.

When open-circuited, the line considered above would be equivalent to a series resonant circuit of only 0·024Ω resistance.

If we arrange to measure the current in the short-circuit of a line as the frequency is varied through the resonance value, we shall obtain a resonance curve from which the Q of the line can be derived by the same relationship as for a "lumped" L.C. circuit, namely, $Q = \frac{f_0}{\Delta f}$, where f_0 is the resonant frequency of the line and Δf is the width of the resonance curve at the points where the current is $\frac{1}{\sqrt{2}}$ of the value at resonance. Since the Q is usually high, it may be necessary to measure the frequency change by a beat method.

Measurement of R.F. Line Properties by Resonant Circuits

When measurements are being made at very high frequencies it will usually be found simpler to use resonant circuit methods

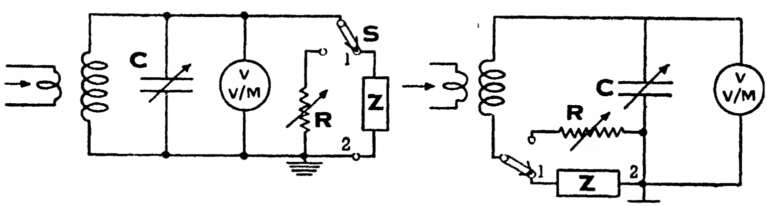


FIG. 125. Measurement of Impedance by Resonant Circuit.

rather than bridges. This is because, provided that the circuit is well arranged, the inevitable stray capacities do little more than reduce the value of tuning capacitance needed, whilst such stray capacities may be extremely troublesome in a bridge network.

A very useful piece of apparatus, therefore, is a resonant circuit consisting of a good coil and condenser, together with a valve voltmeter suitable for high frequencies. For many purposes it is useful to provide two condensers in parallel, one of very small capacitance, and both calibrated. A stable, screened oscillator is required, having sufficient output so that only very weak coupling to the resonant circuit is necessary.

Resonant circuits may be used in a variety of ways. For example, the substitution methods shown in Fig. 125 can be

used, large impedances being measured by the parallel connection and small ones by the series connection. The reactance of Z is found by the change in C necessary to re-tune when it is connected and the resistance by adjusting R until the voltmeter has the same reading for either position of S (tuning to resonance for each reading).

The ordinary " one-watt " composition resistance may be used up to 100 Mc/s in sizes of the order of 1,000Ω with errors less than 5% introduced by the shunt capacitance. The higher the resistance, the more important the shunting capacitance becomes and for a 1 $M\Omega$ resistance might amount to 50%.

An alternative method is to measure the width of the resonance curve at the points where the voltage is $\frac{1}{\sqrt{2}}$ of that at resonance, with and without the impedance under test. If the impedance is a large one, then it will be connected in parallel across the circuit. Having found the reactance of Z by noting the change of C necessary to retune it when Z is connected, the width of the resonance curve is measured, both with and without Z. This will necessitate the use of a very small calibrated condenser in parallel with C. It can be shown that the conductance across the resonant circuit is given by

$$G = \omega \cdot \frac{\Delta C}{2} \mho \qquad . \quad . \quad . \quad . \quad (66)$$

where ΔC is the width of the resonance curve. The conductance of Z is therefore found by the difference between G with and without Z.

In the above discussion Z may, of course, be any kind of impedance. We now wish to see how impedance measurements on lines may be interpreted so as to give Z_o, α, etc. For some of these results it will be necessary to use the equations allowing for losses.

Suppose that we have a sample length of R.F. line and short-circuit the far end. We now set the oscillator to the frequency which will make the line approximately $\frac{\lambda}{4}$ long. If the line is exactly $\frac{\lambda}{4}$ long, then its input impedance will be a pure high resistance. It is therefore possible, by trial and

error, to adjust the frequency until no re-tuning is required when the line is connected. If the oscillator frequency is known and also the length of the sample, then the velocity of propagation along it can be deduced.

By measuring the width of the resonance curve with and without the line, the input conductance of the line is found. From equation (56)

$$G = \frac{1}{Z_0} \tanh \alpha l$$

which can usually be approximated to $\frac{\alpha l}{Z_0}$ since αl will be small.

To measure Z_0 we halve the frequency of the oscillator, so that the line is $\frac{\lambda}{8}$ long. We have seen that, in the case of a loss-free line, the input impedance is then a pure reactance equal to Z_0, being inductive when the line is short-circuited and capacitive when it is open-circuited. This will be a good approximation on an actual R.F. line and hence we can find Z_0. from the change of C necessary when the line is connected. The precise frequency for the $\frac{\lambda}{8}$ condition can be found from the fact that the magnitude of the reactance should be the same whether the line is open or short-circuited.

An example may make the method clearer.

Problem (11). A short-circuited 1·5 metre length of R.F. cable can be connected across a resonant circuit. At a frequency of 40 Mc/s, the cable does not de-tune the circuit.

Without the cable, $\Delta C = 2\ \mu\mu\text{F}$, with the cable it is $2{\cdot}5\ \mu\mu\text{F}$.

At a frequency of 20 Mc/s, connection of the short-circuited cable requires an increase of 80 $\mu\mu$F in the tuning condenser and if the cable is open-circuited the same decrease is required.

Find Z_0, α and the velocity of propagation.

Solution (11). From 20 Mc/s measurements, $Z_0 = \frac{1}{\omega C}$,

or $$Z_0 = \frac{10^{12}}{2\pi \times 20 \times 10^6 \times 80} = 99{\cdot}7\Omega.$$

From 40 Mc/s measurements, difference of $\Delta C = 0{\cdot}5$

$$\begin{aligned} G \text{ of cable} &= 2\pi \times 40 \times 10^6 \times 0{\cdot}25 \times 10^{-12} \\ &= 20\pi \times 10^{-6}\mho. \end{aligned}$$

For $\frac{\lambda}{4}$ length, this $= \frac{\alpha l}{Z}$ or $\alpha = \frac{20\pi \times 10^{-6} \times 99{\cdot}7}{1{\cdot}5}$

$= 4{\cdot}18 \times 10^{-3}$ per metre.

Attenuation in db per km. $= 8{\cdot}686 \times 4{\cdot}18 = 36{\cdot}2$ db.

The wavelength on the cable is $4 \times 1{\cdot}5 = 6$ metres.

Velocity $= 40 \times 10^6 \times 6 = 240 \times 10^6$ metres per second, or 0·8 that in "free space."

Impedance Measurement by Bridge

Bridge methods have the great advantage of being null methods, so that the detector can be made progressively more sensitive. Their accuracy does not depend upon an indicating

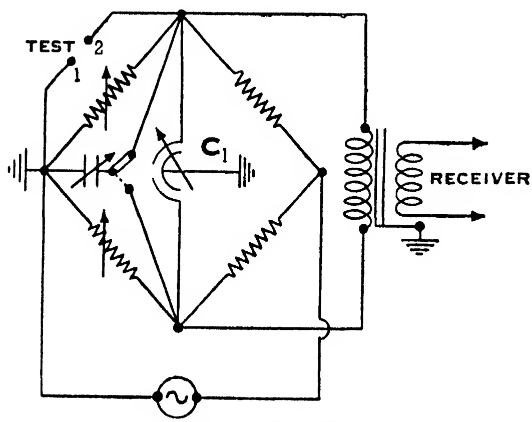


FIG. 126. R.F. Bridge.

instrument following a definite law or remaining constant, or upon the circuit conditions remaining constant while a succession of readings is taken.

We have already mentioned, however, that the design and operation of bridges at high radio frequencies presents difficulties. A number of bridges have been developed which work successfully up to high frequencies and the circuit of one type (developed by the Marconi Co.) which will work up to 40 Mc/s is shown in Fig. 126. A communication type of receiver forms a suitable detector, good screening of oscillator, detector and bridge being essential, with short, screened connections between them.

The bridge is designed to cover measurements over a wide

range of impedance values, a most essential feature when dealing with any type of aerial system, as will be seen in the next chapter. It is of the admittance type, that is to say, it is balanced by obtaining equality of admittances between the arms of the bridge by the addition of a condenser and resistance, in parallel either with the unknown reactance or with the corresponding balancing arm.

Since the unknown reactance may be either inductive or capacitive, the condenser will need to be placed in parallel with the unknown should this be inductive, but will be put in parallel with the variable arm if the unknown is capacitive.

Having found the values of parallel reactance and resistance which give a balance, these are converted to the usual series equivalents by the formula given on p. 183.

The bridge is first balanced with the test terminals on open circuit, by the differential condenser C_1, the setting of which will vary somewhat for different frequencies. The unknown circuit is then connected to the test terminals, and (having first found whether it is inductive or capacitive) the balance point is found by adjusting condenser and resistances.

Experimental Determination of Current and Voltage Distribution along R.F. Lines

It is very instructive to study experimentally the distribution of current and voltage along a line, with different terminations. Also, we have seen that much information can be derived from relative measurements of current. We will now discuss practical apparatus for making the measurements.

For the relative measurements of current, the thermo-junction or one of the modern, permanent forms of crystal detector can be used, in conjunction with a milliammeter. The thermo-junction has the advantage that the relation between heater current and milliammeter reading is definitely a square law. In the case of a crystal this is usually approximately true, but a calibration will be necessary for more accurate work. The crystal has the advantage, however, of being much less likely to be burnt out owing to indiscrete adjustments and of being quicker in response.

The indications of the thermo-junction will depend somewhat upon frequency but the effect is only serious above, say,

100 Mc/s. The increase in heater resistance due to skin effect tends to make the instrument more sensitive for higher frequencies but this may be off-set, particularly with the more sensitive, higher-resistance heaters, by the shunting effect of stray capacitances in parallel with the heater wire.

The crystal tends to become less sensitive with increasing frequency, principally because the capacitance between point contact and crystal allows of an unrectified R.F. current but the effect is small until centimetric wavelengths are reached.

In making R.F. line measurements we seldom wish to

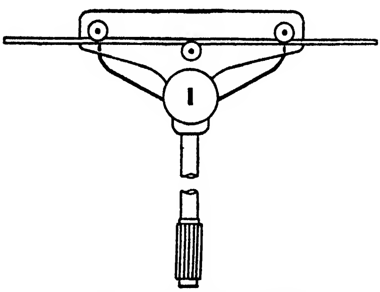


FIG. 127. Line Current Indicator.

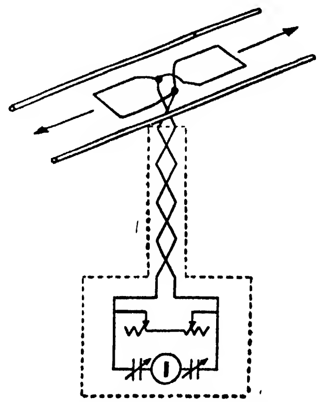


FIG. 128. Line Current Indicator.

compare currents at different frequencies or to determine their actual values.

If a line is being set up merely for the purpose of studying the current distribution along it, we shall naturally use a parallel-wire type, on account of its accessibility. A good arrangement is to erect the line about 2 metres above the ground floor, so that adjustments are made from beneath. The presence of the observer has then very little disturbing effect. If indoors, the line should be about 2 metres from a parallel wall and parallel wires, conduits or pipes should be avoided as, otherwise, it will usually be found that the currents in the two wires are quite different.

A simple instrument, which enables the current to be read in either line, is shown in Fig. 127. The line will need to be well cleaned before use.

A rather better type of instrument is shown in Fig. 128, which is designed to be free from all contacts, but at the same time picks up energy only through electromagnetic induction. This is accomplished by designing the coil in two halves, cross-connected, the output from the coils being connected through a screened feeder to the tuned circuit and indicating meter. The whole instrument is carried on an insulating slider (not shown) which locates the coil exactly midway between the lines, and it has been found that very accurate results can be obtained.

An alternative instrument, which has been found very satisfactory, uses only one turn but has an electrostatic screen of wires above and below it. The connections from the coil are led through a screening tube into a screening can containing a crystal detector and milliammeter, no tuning being employed.

The current distribution is all that is usually required but if it is desired to obtain the voltage distribution this can be done by using a short wire connected to a crystal detector.

The current distribution in a concentric line cannot be obtained, unless the line is specially arranged for the purpose. In centimetre-wave test equipment it is usual to provide a length of concentric line with a slot, along which a detecting device can travel. This may be either a short, straight wire forming a probe, or a small loop, a crystal detector and milliammeter being used. The probe will trace out the voltage distribution and the loop the current distribution.

Measurement of Impedances by Means of Lines

When wishing to measure impedances by frequencies such as 300 Mc/s, it is desirable to use a line rather than a resonant-circuit method.

We have already seen (see page 195) that an impedance placed across a line of known Z_0 can be determined if the ratio of minimum to maximum currents and the position of the current minimum with respect to the load is known. Hence a length of line carrying a current indicator moving along a graduated scale can be used to measure impedances.

This method can have an accuracy of about 2% when measuring impedances not too different from Z_0 but if the impedances are larger than about $10Z_0$, or smaller than about

$0{\cdot}1Z_0$, then the reflected wave is nearly equal to the incident and changes but little for a change in load. The method is, therefore, insensitive under such conditions.

An alternative method, very suitable for a wide range of impedances, is illustrated in principle by Fig. 129. A length of line has a movable short-circuiting bridge, B, carrying a current indicator, TJ and G. A short-circuiting disc is first put across the line, in place of Z_x, and the variation of G noted

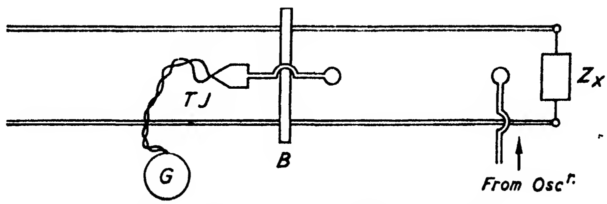


FIG. 129. Measuring Impedances by Line.

as B is moved. The current will be a maximum when the used portion of the line is $\frac{\lambda}{2}$ and the width of the resonance curve obtained is related to the Q of the line.

The unknown impedance, Z_x, is then connected and the procedure repeated. From the new position of B and the width of the resonance curve, the resistance and reactance of Z_x can be determined. The reader is referred to the original papers for the derivation of the necessary formulæ. A number of variations of this method have been published.

The accuracy obtainable from both the above methods will depend considerably upon good mechanical design and workmanship. Concentric lines will be used for measuring unbalanced impedances and parallel-wire lines (usually in a cylindrical screen) for balanced impedances.

Selected References

(1) FLEMING. *The Propagation of Electric Currents.* Constable.
(2) MALLETT. *Telegraphy and Telephony.* Chapman and Hall.
(3) HARTSHORN. *Radio Frequency Measurements.* Chapman and Hall.
(4) WILLIS JACKSON. *High Frequency Transmission Lines.* Methuen.
(5) GREEN. "Short Wave Aerial Systems." *W.E.*, June, 1928.

(6) CORK AND PAWSEY. "Long Feeders for Transmitting Wide Sidebands." *J.I.E.E.*, vol. 84. 1939.

(7) HAYES AND MACLARTY. "Empire Broadcasting Station." *J.I.E.E.*, vol. 85, 1939.

(8) WILLIS JACKSON AND FORSYTH. "Polythene as a High-Frequency Dielectric." *J.I.E.E.*, vol. 92. March, 1945.

(9) McLEAN AND BOLT. "R.F. Open-Wire Transmission Lines." *J.I.E.E.*, vol. 93. May, 1946.

(10) GENT AND WALLIS. "Impedance Matching by Tapered Transmission Lines." *J.I.E.E.*, *R.L.C.*

(11) GREEN. "Quarter Wave Networks." *Mar. Rev.* October–December, 1949.

CHAPTER VII

WAVE GUIDES

It was shown by Rayleigh in 1897 that it is possible for electromagnetic waves to be propagated in a dielectric enclosed by conducting walls (without a second conductor inside) but the cross-section of the enclosure must have dimensions comparable with the wavelength. Hence only centimetre waves can economically be used and it has only become possible in recent years to generate the very high frequencies required.

It is now known both theoretically and experimentally that the attenuation at very high frequencies can be considerably less in a wave guide than in a concentric cable. Less conducting material is required and no solid dielectric. Hence wave guides are becoming of increasing practical importance for linking very high frequency oscillators to aerials, etc., and may be used in the future for " line " transmission, employing a very high frequency carrier modulated by a large number of telephone channels.

If conductor losses are neglected, then the wave in a concentric cable which has a diameter large compared with the wavelength is of the same type as in free space. The electric and magnetic fields are perpendicular to each other and to the direction of propagation. The wave is entirely transverse.

The only waves which can travel along a wave guide, however, have *either* a magnetic or electric component *along* the direction of propagation, beside transverse components, even if the walls are perfect conductors.

Most British writers have classified waves in guides by the longitudinal component—thus an H-wave has a component of magnetic field along the direction of propagation. Most American writers, on the other hand, describe a wave by the purely transverse component and thus the H-wave previously mentioned would be called a TE (transverse electric) wave. The American Institute of Radio Engineers has recently published a suggested set of standard definitions based on this system.

The treatment of the subject in this chapter will be mainly descriptive and references to more theoretical and complete treatments will be found at the end of the chapter.

A number of H and E waves are possible, but in the early part of this chapter we shall deal with only one important type, to prevent confusion.

Two "Intersecting" Plane Waves

As a preliminary to considering waves travelling through a rectangular enclosure, let us consider two plane, electromagnetic waves of equal amplitudes, as shown in Fig. 130, travelling in

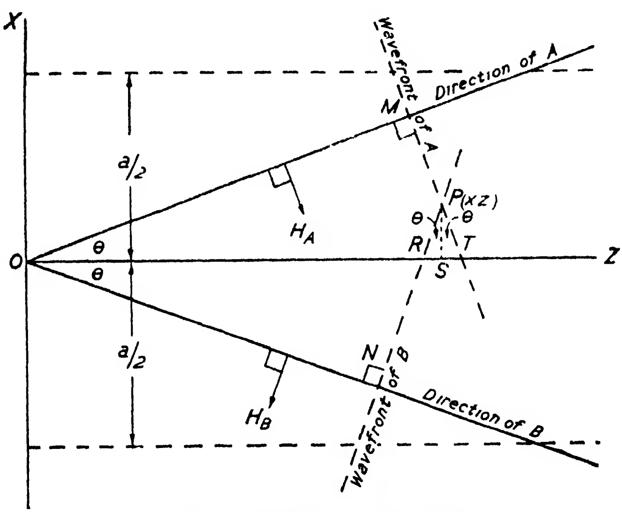


Fig. 130. Intersecting Plane Waves.

space along paths which intersect at an angle 2θ. For the moment we will deal only with the electric fields, which will both be in a direction perpendicular to the plane of the paper. Let these be varying sinusoidally with time and have a maximum value E. We will choose the origin of our X, Y and Z co-ordinates at a point where both waves have the same phase and hence the electric fields of both may be written $E \cos \omega t$.

Now consider point P. The wavefront of A has advanced a distance OM which is $OT \cos \theta = (OS + ST) \cos \theta$

$$= z \cos \theta + x \tan \theta \cos \theta$$
$$= z \cos \theta + x \sin \theta$$

so that $E_A = E \cos\left[\omega t - \frac{\omega}{c}(z \cos\theta + x \sin\theta)\right]$

where c is the velocity of the component waves.

The wavefront of B has advanced ON, which is $z \cos\theta - x \sin\theta$ and hence the resultant electric field

$$E_R = E \cos\left[\omega t - \frac{\omega}{c}(z \cos\theta + x \sin\theta)\right]$$

$$+ E \cos\left[\omega t - \frac{\omega}{c}(z \cos\theta - x \sin\theta)\right]$$

$$= 2E \cos\left(\frac{\omega}{c} x \sin\theta\right) \cos\left\{\omega t - \frac{\omega}{c} z \cos\theta\right\}. \quad . \ (1)$$

Only z appears in the term which specifies the phase of the resultant wave, which is therefore travelling in the z direction, the velocity being $\frac{c}{\cos\theta}$.

So far we have considered only the resultant electric field but we will now see how the component magnetic fields, which lie in the XZ plane, add to form a resultant magnetic field.

Resolving the magnetic fields along the X and Z axes, we have, for the point P

$$H_{XA} = -H \cos\theta\left[\cos\left\{\omega t - \frac{\omega}{c}(z \cos\theta + x \sin\theta)\right\}\right]$$

$$H_{XB} = -H \cos\theta\left[\cos\left\{\omega t - \frac{\omega}{c}(z \cos\theta - x \sin\theta)\right\}\right]$$

$$H_{ZA} = -H \sin\theta\left[\cos\left\{\omega t - \frac{\omega}{c}(z \cos\theta + x \sin\theta)\right\}\right]$$

$$H_{ZB} = -H \sin\theta\left[\cos\left\{\omega t - \frac{\omega}{c}(z \cos\theta - x \sin\theta)\right\}\right]$$

Hence

$$H_{ZR} = 2H \sin\theta\left[\sin\left(\omega t - \frac{\omega}{c} z \cos\theta\right) \sin\left(\frac{\omega}{c} x \sin\theta\right)\right] \quad (2)$$

$$H_{XR} = -2\mathrm{H} \cos\theta\left[\cos\left(\frac{\omega}{c} x \sin\theta\right) \cos\left\{\omega t - \frac{\omega}{c} z \cos\theta\right\}\right] \quad (3)$$

It will be seen that the resultant wave differs in character

from the component waves because there is a component of H along the direction of propagation. This longitudinal component is in time quadrature with the electric field, though the transverse magnetic field is in time phase with the electric field, as in the component waves.

The velocity of both H_x and H_z is given by

$$c/\cos\theta \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

as in the case of the electric field.

Transmission between Parallel Conducting Planes

From equation (1) we see that the resultant electric field varies with x in accordance with the term $\cos\left(\frac{\omega}{c}\,x\sin\theta\right)$ which will become zero if $\frac{\omega x}{c}\sin\theta = \pm\frac{\pi}{2}$.

Hence E will always be zero along two planes parallel to the Z axis and distant from it by

$$\frac{a}{2} = \frac{\pi}{2}\cdot\frac{c}{\omega\sin\theta} = \frac{c}{4f\sin\theta}.$$

Consequently, if perfectly-conducting walls, perpendicular to the plane of the paper (see Fig. 130) are placed at a distance $\frac{a}{2}$ from the axis the electric field between the planes need not be modified, since the condition that there can be no electric field along the surface of a perfect conductor is satisfied.

Suppose now that the conducting planes have been fixed at a distance apart a then

$$\sin\theta = \frac{c}{2fa} \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

so that, as the frequency is lowered, θ becomes larger. The limiting condition is when $\theta = \frac{\pi}{2}$ making

$$f_0 = \frac{c}{2a} \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

where f_0 is the lowest frequency at which the resultant wave travelling in the z direction can be produced.

It is useful to re-write (5) in the form

$$\lambda_0 = 2a \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

where λ_0 is the longest wavelength of the component waves (that is, corresponding to a velocity c) and *not* the wavelength of the resultant wave.

Transmission down a Rectangular Wave Guide

Suppose that we form a hollow box by placing two more conducting planes parallel to the XZ plane and at a distance $\frac{b}{2}$ above and below it. Then whatever the distance b is, the boundary conditions are satisfied because the electric field is normal to the conducting planes.

We see, therefore, that a wave can travel down a rectangular guide provided that the frequency is sufficiently high, the guide behaving as a high-pass filter. The cut-off frequency is

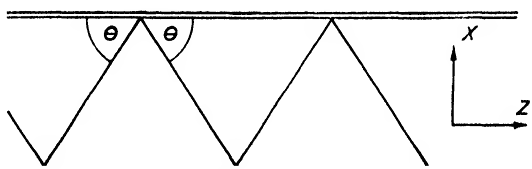


FIG. 131. Transmission Between Parallel Planes.

determined by the distance apart of the two walls which are parallel to the electric field. In the ideal case of perfectly conducting walls there will be no attenuation but since currents flow in the walls there will be loss in the practical case. Transmission down the guide is possible whatever the distance apart of the other two walls but the attenuation is affected by this spacing, being a minimum for a nearly square guide.

The type of wave we have so far considered is characterised by having an H component along the direction of propagation and in time quadrature with the transverse H and E components.

We have so far considered the transmission as the resultant of two independent waves intersecting. It is not necessary, however, to launch two independent waves into the guide because one is formed from the other by reflection at the walls (Fig. 131). Thus the whole guide is filled, by reflections,

with the component waves we have postulated as necessary to initiate transmission down the guide.

The facts that we have deduced from the above argument can be obtained more directly and rigorously by putting the boundary conditions into Maxwell's equations and finding possible wave solutions without reference to the means by which the waves may be produced. This method has been

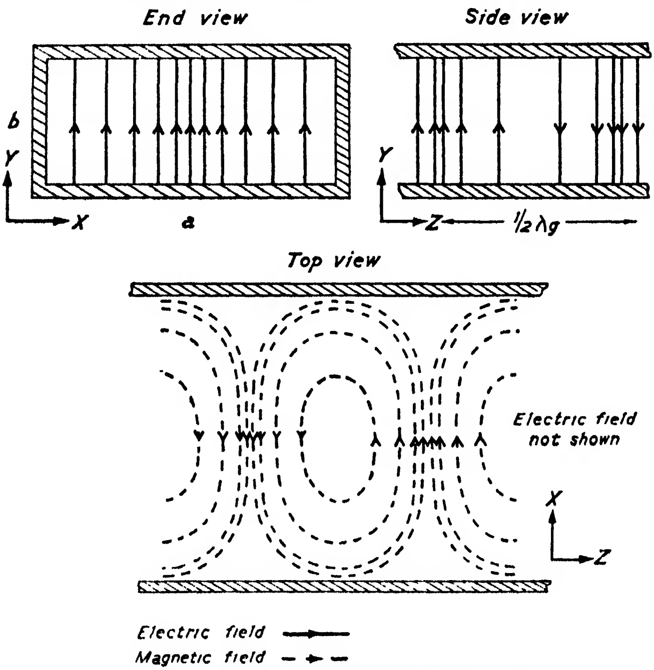


FIG. 132. $H_{1,0}$ Wave in Rectangular Guide.

followed in the classical papers of Carson, Southworth and others, and will be found set forth in several text-books.

More recently Moullin has worked out the field which would exist down a guide from a given current in a wire inserted in the guide.

The field distribution for the type of wave we have been considering is shown in Fig. 132. The distribution shown in side and top views will travel down the guide at the velocity $\dfrac{c}{\cos\theta}$.

Phase and Group Velocities

We can see from (4) and (5) that the velocity of the resultant wave varies with frequency, unlike that of a wave in "free space." Whenever this is the case (a wave travelling through the ionosphere is an example that we have already dealt with) then we have to distinguish between *phase* and *group* velocity.

If we measure the distance between the nearest two points at which the phase is the same at the same instant (that is, a wavelength) then we can find the velocity with which phase changes are propagated. In a length of wave guide closed at the ends, for example, there would be large stationary waves built up and the positions of two adjacent minima could be found, in a similar way to that discussed for a line. These would be $\frac{\lambda_g}{2}$ apart and, if the frequency of the source is f, then the phase velocity would be $v_{ph} = f\lambda_g$. It will be seen that it is v_{ph} which has been specified by (4) and that it is higher than c, the velocity of the component waves.

$$\text{From (5)}\quad \cos\theta = \sqrt{1 - \sin^2\theta} = \sqrt{1 - \frac{c^2}{4f^2a^2}}$$

$$= \frac{c}{\omega}\sqrt{\frac{\omega^2}{c^2} - \frac{\pi^2}{a^2}}$$

and hence

$$v_{ph} = \frac{\omega}{\sqrt{\frac{\omega^2}{c^2} - \frac{\pi^2}{a^2}}} \quad . \quad . \quad . \quad . \quad (8)$$

which at the cut-off frequency becomes (from (6)) infinite.

Suppose now that we initiate a short pulse of high-frequency waves at one end of the wave guide and could measure the time taken before it reached the other end. This measurement would give us the group velocity, v_{gr}, and v_{gr} is different from v_{ph} for any medium in which the phase velocity depends upon frequency. This is because any wave lasting a finite time comprises a band of frequencies and these travel at different velocities.

Consider, as an example, the two component waves of Fig. 133 having wavelengths of λ and $\lambda + \delta\lambda$ and phase velocities of v

and $v + \delta v$ respectively. Due to the different velocities, the crest l' will in time catch up to the crest l and hence the maximum value of the group now corresponds to ll'. It

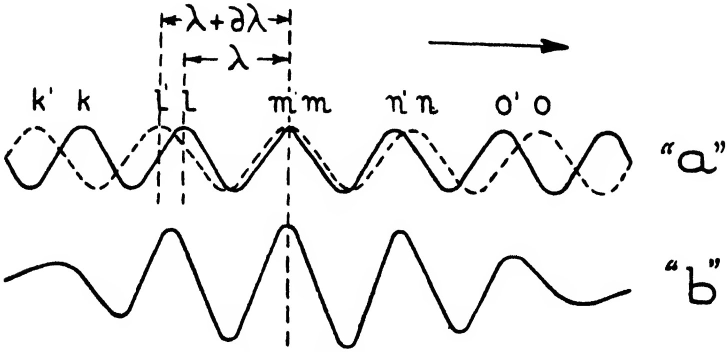


FIG. 133. Illustrating Group Velocity.

will be seen that the group has fallen back with respect to individual waves ; in other words, group velocity is less than phase velocity.

The relative velocity of l' with respect to l is δv and hence it will take a time $\frac{\delta\lambda}{\delta v}$ to catch up on l. When this has occurred, the group will have moved back λ relative to individual wave crests. Hence

$$v_g = v - \lambda \frac{\delta v}{\delta \lambda}.$$

It can be shown that the variation of phase velocity with frequency, both in a wave guide and in the ionosphere, leads to the relation

$$v_{ph}\, v_{gr} = c^2$$

If our pulse involves a large band of frequencies, then its envelope will change its shape considerably during transmission and the group velocity is then somewhat indefinite. For example, in our wave guide experiment we might initiate a pulse which rose to its full value very rapidly, so that the instant of starting it off was very definite. After transmission down the guide, however, it would become rounded and it would not be easy to decide the instant of its arrival.

It should be realised that v_{gr} is the velocity with which energy is being transmitted down the guide. The only way in

which we can measure this velocity is by a short pulse, in which case it is clear that the " packet " of energy in the pulse has travelled down the guide at the group velocity. If we keep the transmission going for a considerable time, however, it is still true that the energy is travelling at the group velocity whilst the phase changes are travelling at the phase velocity.

The wavelength, λ_g, measured in a guide, is clearly longer than the wavelength in space corresponding to the same frequency, because v_{ph} is higher.

An example may help to fix our ideas.

Example (1). In a certain wave guide stationary waves exist, due to reflection from the ends, and the distance between successive points of maximum electric field is found by sliding a probe in a slot along the guide and is 12·5 cm. The guide supplies a coaxial line (with air dielectric) on which the distance between successive maxima is found to be 7·5 cm. Estimate (*a*) the frequency of the source supplying the guide, (*b*) the phase and group velocities within the guide, (*c*) the lowest frequency which could be transmitted down the guide.

Solution (1). If losses in the concentric line are assumed to be negligible, then the phase (and group) velocity will be

$$c = 3 \times 10^{10} \text{ cm/sec.} \quad \lambda/2 = 7{\cdot}5 \text{ cm.}$$

$$f = \frac{c}{\lambda} = \frac{3 \times 10^{10}}{15} = 2{,}000 \text{ Mc/s.}$$

The wavelength λ_g in the guide is $2 \times 12{\cdot}5$ cm and hence phase velocity $= f\lambda_g$ or

$$v_{ph} = 2 \times 10^9 \times 25 = 5 \times 10^{10} \text{ cm/sec.}$$

Since $\quad v_{ph}\, v_{gr} = c^2. \quad v_{gr} = \dfrac{9 \times 10^{20}}{5 \times 10^{20}} = 1{\cdot}8 \times 10^{10}$ cm/sec.

From equation 8 we find $a = 75/8$ cm.

The lowest frequency which can be transmitted will have a " free-space " λ of $2a = 75/4$ cm

and $f = 1{,}600$ Mc/s.

Attenuation of * $H_{1,0}$ ($TE_{1,0}$) Wave in Rectangular Guide

From a knowledge of the field distribution the currents in the walls can be calculated and the high-frequency resistance of the walls can also be calculated, so that the conductor losses may be known. Clearly, at the very high frequencies which

* The type of wave we have so far been discussing is only one of many possible types and the suffixes (the significance of which are explained later) specify it.

will be used, only a thin skin of conductor will carry current and therefore the thickness of the walls will not affect the result.

In Fig. 134 the variation of attenuation with frequency in guides having different a/b ratios but all of the same perimeter (28 cm) and all using the same amount of material (copper) is

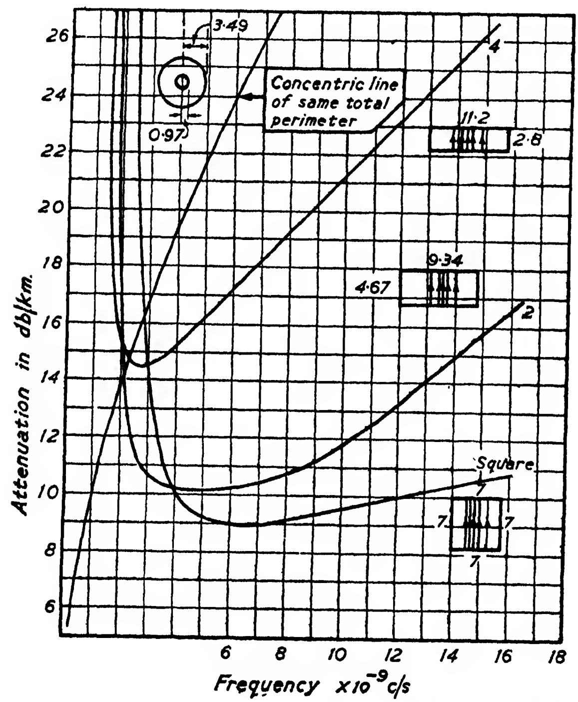


FIG. 134. Attenuation-Frequency Curves for Rectangular Wave Guides and Concentric Line.

shown. For comparison, the attenuation (due only to conductor loss) in a concentric line using the same amount of copper and having the best ratio of outer to inner conductor (3·65) is shown.

It will be seen that a square guide has less attenuation for a given periphery than other rectangular guides. On the other hand, by making a several times greater than b, we can get lower frequencies through the guide for the use of the same amount of material.

At very high frequencies the curves show that the guide introduces much less attenuation than the concentric line and it should be remembered that there will be some supporting insulators in the concentric line, which will introduce serious additional losses at U.H.F.

Resonant Wave Guides

If a guide is closed by a perfectly-conducting plate, then a reflected wave must be formed so that there is no resultant electric field over the plate. Thus stationary waves are set up and if the length of the guide is suitably adjusted, these

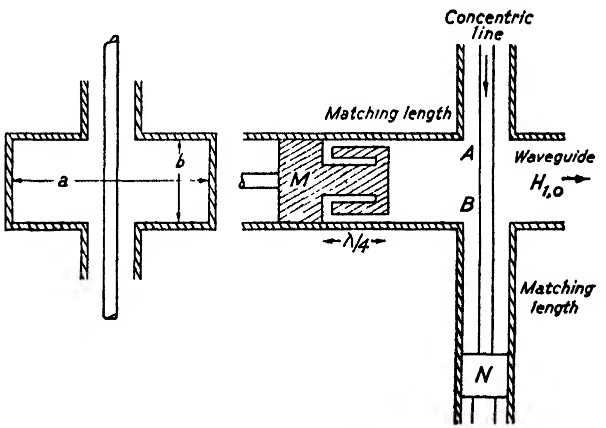


FIG. 135. Production of $H_{1,0}$ Wave in a Guide.

can become large. The behaviour is, therefore, closely analogous to that of the short-circuited line.

If both ends of the guide are closed and made $\frac{\lambda_g}{2}$ apart, then the guide will be in resonance, the electric field being always zero at each end and a maximum half-way along the guide. Note that the length for resonance is one-half of the wavelength *in the guide* which is, of course, longer than the wavelength in free space for the same frequency.

A suitable adjustable piston for closing the end of the guide is included in Fig. 135. The half-wave trap (folded back on itself) being shorted at one end presents a short across the guide

at the other end. Hence the current in the actual sliding contact $\left(\frac{\lambda}{4}\text{ away}\right)$ is very small.

If a detector is fixed in position in the guide and the piston varied through the $\frac{\lambda_g}{2}$ position, a very sharp resonance curve is obtained, from which it is evident that the effective Q of such a resonant guide is very high. The Q may be several thousands and can be calculated, since there are no losses in insulators to consider.

Production of $H_{1,0}$ ($TE_{1,0}$) wave in Guide

If we put a wire across a guide half-way along a (as in Fig. 135), then when this wire carries current, the electric and magnetic fields produced will be of the same general shape as the field pattern for the $H_{1,0}$ wave and hence this type of wave will be produced in the guide.

A similar arrangement would be used to extract energy from a guide—to feed a concentric line, for example.

In order to make the transfer efficiently, with minimum losses in guide and concentric cable, it will be necessary to match the impedances at the junction. By adjusting M we can ensure that radiation to the left from AB, reflected from M, is in the right phase on arrival back at AB to reinforce radiation going out to the right. By adjusting N, the impedance placed across the end of the main concentric cable can be made equal to its characteristic impedance.

Other Orders of H Waves (TE Waves) in Rectangular Guides

We have so far discussed only one type of wave in order to prevent confusion, but many other field patterns will also

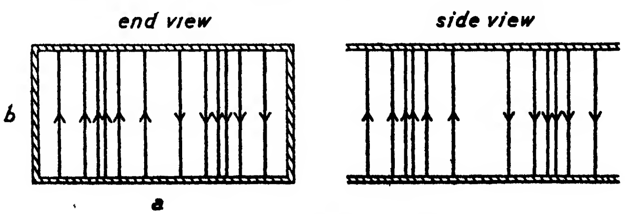


FIG. 136. $H_{2,0}$ Wave in Rectangular Guide.

satisfy the boundary conditions and therefore many other types of waves can be propagated. These will only be discussed briefly and some typical field patterns illustrated.

For example, if the frequency is sufficiently high, then we can have two maxima of the electric field along a whilst still having zero electric field at the sides, as seen in Fig. 136.

The order of a wave in a rectangular guide is specified by two suffixes m and n, which refer to the number of half-sines in electric field distribution along a and b respectively. We now see the meaning of the suffixes we have applied to the $H_{1,0}$ we have been discussing.

The general expression for the cut-off is given by

$$f_o = \frac{c}{2}\sqrt{\frac{m^2}{a^2} + \frac{n^2}{b^2}}$$

and it will be seen that the $H_{1,0}$ wave has the lowest cut-off frequency of any H type.

The attenuation is also least for this type and, therefore, the $H_{1,0}$ wave is of great practical importance.

E Waves in Rectangular Guides

The electric lines must still end normally on the walls of the guide but must be so arranged that there is a longitudinal

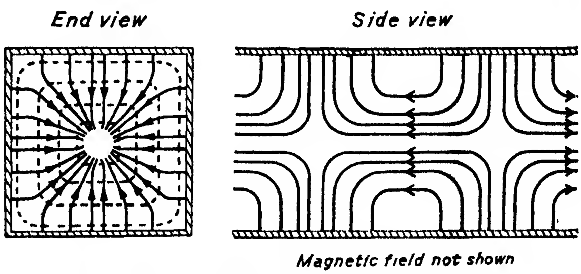


FIG. 137. $E_{1,1}$ Wave in Rectangular Guide.

component. The simplest pattern which is possible is that of the $E_{1,1}$ wave shown in Fig. 137. The cut-off frequency is given by the same expression as for H waves. In the case of E waves, neither m or n can be zero. Since the electric field parallel to the side must be zero at that side, it would be zero over the whole cross-section if m or n were zero. It follows that the $H_{1,0}$ wave has the lowest cut-off frequency of *any* type.

Circular Guides

Guides of circular cross-section will clearly be convenient for many purposes and it therefore becomes necessary to study the possible types of waves in them.

A possible H wave (TE wave) is shown in Fig. 138. The electric lines can be circular because the electric field is zero

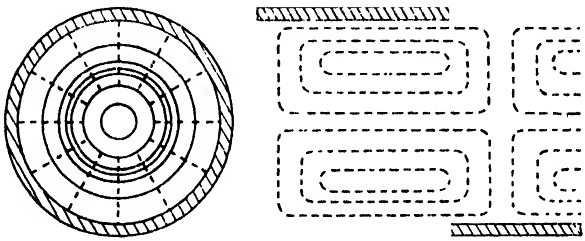


FIG. 138. $H_{0,1}$ Wave in Circular Guide.

at the surface of the conductor and therefore the requirement that there shall be no tangential electric force is fulfilled.

The order of the wave in a circular guide is arrived at in the following way. In the case of an H wave when the first suffix

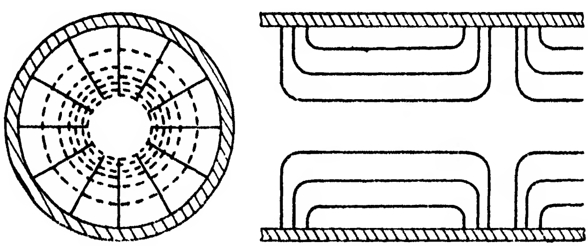


FIG. 139. $E_{0,1}$ Wave in Circular Guide.

is zero the lines of electric field are circular so that we meet no maxima or minima if we traverse a concentric circle in the guide. The second suffix then gives the number of concentric cylinders (including the conducting boundary) at which the electric field is zero.

When the first suffix of an E-wave is zero, the magnetic lines are circular and the second suffix shows at how many cylinders (including the conducting boundary) the electric lines are normal. A reference to Figs. 138 and 139 should make these

terms clear, it being realised that, in the case of the E-wave, the electric lines are only exactly normal to the bounding cylinder.

If the first suffix is not zero, then neither the electric or magnetic field is zero. For both H and E-waves, the first suffix gives the number of axial planes and the second the number of cylinders (including the conducting boundary) to which the electric lines are normal.

The $H_{1,1}$ wave is of interest because this type has the lowest cut-off frequency.

The cut-off frequency for circular guides can be found from

$$\omega_c = \frac{c\gamma_n}{a}$$

where c is the velocity in free space, a is the radius of the guide and γ_n is a constant which is different for each type of wave.

Values of γ_n for some types of wave are given below :

	$E_{0,1}$	$E_{1,1}$	$H_{0,1}$	$H_{1,1}$
γ_n	2·4	3·8	3·8	1·8

It will be seen that the $H_{1,1}$ type allows the lowest frequencies to be transmitted along a guide of given size.

Attenuation in Circular Guides

Fig. 140 shows the attenuation of various waves in a copper circular guide compared with the attenuation due to conductor

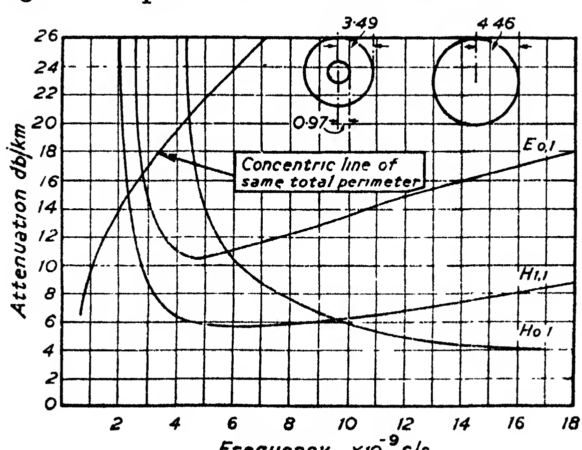


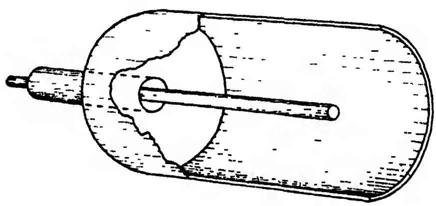
FIG. 140. Comparison Between Attenuations of E and H Waves in Circular Guide of 28 cm Periphery.

loss in a concentric line using the same material and with an inner conductor having a diameter of the best ratio (3·6) to that of the outer conductor. The curves show the superior performance of the guide at the higher frequencies.

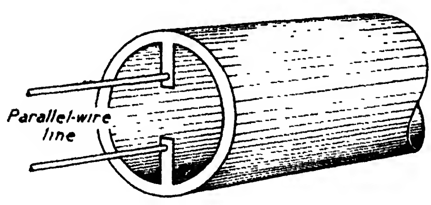
The attenuation of the $H_{0,1}$ wave is unique in that it decreases without limit as the frequency is raised. This is because the current in the conducting cylinder decreases with increasing frequency. Theory shows that this is not true if the cross-section is elliptical and the effect of quite a small deformation of the cylinder will be to increase the attenuation and probably to cause a change to another type of wave.

Production of Waves in Circular Guides

The production of an $E_{0,1}$ wave is easily accomplished by using a short axial conductor. If the energy is being supplied



(a) *Excitation of $E_{0,1}$ wave from Concentric Line*



(b) *Excitation of $H_{1,1}$ wave from Parallel-wire Line*

FIG. 141. Excitation of Waves in Cylindrical Guides.

through a concentric cable the arrangement becomes as shown in Fig. 141a.

The H types require a rather more elaborate arrangement because the electric field is now circular. One arrangement is shown in Fig. 141b.

Energy can be transferred to a guide from a concentric cable, or from one guide to another, through a slot. A narrow slot lying along the electric field has little effect but if the slot is transverse then much of the energy is transferred through the slot. Slots as radiators are further discussed in the next chapter.

Matching Wave Guides

When it is desired to place a reactance across a guide an iris can be used as the equivalent of the stub which would be used on a line. Three types for use in a rectangular guide carrying

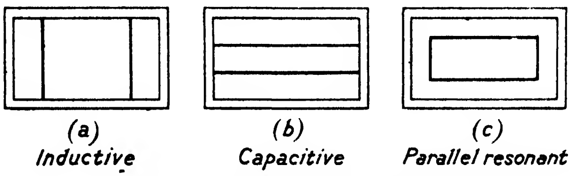


FIG. 142. Matching Irises.

an $H_{1,0}$ wave, are shown in Fig. 142. In each case the iris will be only a small fraction of a wavelength long in the direction of propagation, so that it can be regarded as a concentrated reactance.

Wave Guides as Attenuators

If a frequency lower than the cut-off frequency is applied to a guide, the attenuation does not become infinite but may be 1 db per mm. This high value is not dependent upon the material of the walls, as it is not due to losses in them. It is also nearly independent of frequency, as long as this is well below cut-off, and can be calculated from the dimensions of the guide.

In ultra-high frequency signal generators, a wave guide working in this way has been found a suitable means for obtaining the required known attenuation. A circular guide is usually used and a source arranged at one end. A sliding piston carries a pick-up arrangement and hence a variable, known attenuation can be obtained by moving the piston, the arrangement being known as a piston attenuator. The calculation of attenuation does not hold when the pick-up is too near the source and hence there is an initial attenuation

which may be 30 db. Care has to be taken that only one type of wave is launched into the guide, because the attenuation is different for different modes.

Dielectric Wave Guides*

It is possible to guide electromagnetic waves along a cylinder or rectangle of dielectric without any surrounding conductor. For the guiding to be effective it will, naturally, be necessary for the dielectric constant of the material used to be much greater than that of the surroundings.

If a wave is initiated within the cylinder, then there will be reflection at the boundary between the dielectrics and conditions will therefore be somewhat similar to those in a metallic guide. though there will be appreciable energy in refracted waves which will pass out into the surrounding dielectric.

The theory of dielectric guides has been studied and experimental confirmation obtained, notably by Southworth, using a cylinder of water as the guide.

The dielectric wave guide has found little application as yet. The attenuation in solid dielectrics having high dielectric constants is high and there would appear to be no practical or economical advantages of this type over the metallic guide with air as dielectric.

Selected References

(1) LAMONT. *Wave Guides.* Methuen.
(2) SLATER. *Microwave Transmission.* McGraw-Hill.
(3) BARLOW. *Micro-waves and Wave Guides.* Constable.
(4) SOUTHWORTH. "Hyper-frequency Wave Guides." *B.S.T.J.*, 1936, p. 284.
(5) BARROW. "Transmission of E.M. Waves in Hollow Tubes." *P.I.R.E.*, vol. 24. October, 1936.
(6) CLAVIER. "Theoretical Relationships of Dielectric Guides." *Elec. Com.*, January, 1939.
(7) CLAVIER AND ALTOVSKY. "Experimental Researches on Dielectric Guides." *Elec. Com.*, July, 1939.
(8) KEMP. "Wave Guides in Electrical Communication." *J.I.E.E.*, vol. 90. September, 1943.
(9) MOULLIN. "Propagation of Electric Waves in a Rectangular Wave Guide." *J.I.E.E.*, vol. 92. March, 1945.
(10) PRYCE. "Wave Guides." *J.I.E.E., R.L.C.* (and other *R.L.C.* papers).

* The hollow-conductor wave guide was at one time termed a dielectric guide by some workers.

AERIALS

An aerial to be used for long waves will always be less than $\frac{\lambda}{4}$ high and the wire is always earthed, in order to utilise its height to the best advantage. The usual long-wave aerial consists of a vertical portion which supplies the useful, vertically-polarised radiation and a horizontal portion, or "roof." The capacitance of this serves to increase the current in the vertical portion and make it more uniform, thereby increasing the radiation from it.

Owing to the height being only a fraction of the wavelength the radiation resistance is small and it is necessary to keep the total resistance to a low value if a reasonable proportion of the power put into the aerial is to be radiated. For this reason, and also because we are concerned with radiation along the earth's surface, it is essential to provide a very low-resistance earth system. This is done by burying a network of wires in the ground or by erecting an earth screen, that is, a network of wires above the earth's surface, in addition to an earthing point.

Even with the roof, the aerial length will normally be less than $\frac{\lambda}{4}$ and hence it can be tuned by adding inductance between the base of the aerial and earth. Thus a large tuning range can be covered, although, as the dimensions of the aerial become a smaller and smaller proportion of the wavelength being used, the radiated power for a given input power will decrease.

Considering now aerials for short waves, it will be evident that physically they could be and often are many wavelengths long. The roof can be dispensed with, because if the aerial is a quarter-wave or more long there will be no difficulty in producing a large stationary wave of current without having an unduly large voltage at the high potential end. One of the advantages of short waves is that a small and simple aerial will accept a large output from a transmitter and radiate a large proportion of it.

Since the short wave aerial is such a much more efficient radiator and also because we are usually concerned with radiation leaving the earth's surface at an angle, the provision of elaborate earthing systems is not necessary.

In many cases the aerial will not be earthed but will be at a height above the earth comparable with the wavelength, and it may be arranged to produce radiation polarised in any direction, both vertical and horizontal polarisation being used.

The tuning of a short wave aerial is a more complicated matter than in the case of long waves, because its behaviour is subject to sudden disconcerting changes as the applied frequency is changed, and hence considerable difficulty may be experienced in "fitting" the aerial to a circuit unless proper precautions are taken.

In general, we have two distinct types of short wave aerials—those which are designed to give the greatest possible efficiency at one given wavelength and are not usually suitable for covering any wave range, and those which cover a wave range and sacrifice efficiency to achieve this.

Distribution of Current, etc.

Before discussing the radiating properties of the various types of aerials, we will consider the behaviour of an aerial as a circuit. In particular, we require to know the current distribution along it, and its impedance measured between the points at which it is to be fed.

In the case of a vertical aerial this will usually be between the base end and earth, and in the case of a horizontal aerial usually at its mid-point. A knowledge of the impedance at the feed point is necessary in order that we may arrange suitable coupling for transferring energy to or from the aerial when it is connected to transmitter or receiver.

Although electromagnetic waves are produced whenever a varying current flows along a conductor, and certain short wave aerial systems operate by virtue of a travelling wave along a correctly terminated wire, radiation is usually associated with a system in which stationary waves are present, and we will consider the operation of aerials working under such conditions. In order to radiate at all efficiently it is necessary that the dimensions of the circuit should be com-

parable with the wavelength, and the simplest method of getting a large current into such a circuit is to produce stationary waves.

Radiation is greatest from those parts of a circuit where the varying current is greatest (unless influenced by adjacent conductors), and if the length of the system is long compared with the wavelength several modes of oscillation are possible.

Whatever the length or form of aerial, it is clear there can be no current at the end remote from the earth, but when a vertical aerial is considered with one end near earth it may be possible, by suitably arranging a circuit between the lower end

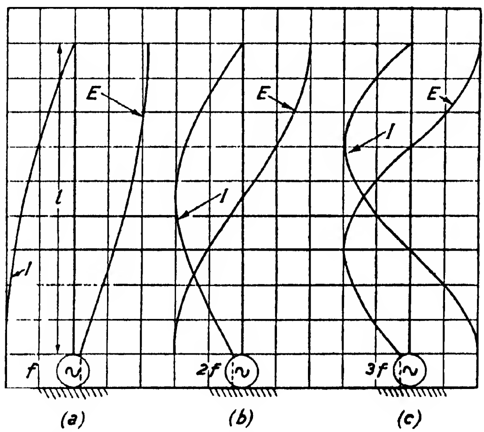


FIG. 143. Voltage and Current Distribution in Aerials.

of the aerial and earth, to produce stationary waves such that either a voltage or a current node is produced at the lower end.

Consider a vertical wire of length l, fed from a generator of frequency f, as shown in Fig. 143. If we induce current in it such that the wire is approximately one-quarter of the wavelength in free space corresponding to f, then by a small variation of the length of wire (or of the applied frequency) we shall obtain a resonance at which the current is a maximum at the base of the wire. This resonance point is called the natural wavelength, and the distribution of current and voltage will be as shown in Fig. 143a, from which we can observe that the wire is carrying a $\frac{\lambda}{4}$ stationary wave. The actual length of the wire is less

than a quarter of the wavelength, and more nearly $\frac{\lambda}{4 \cdot 5}$, because the capacity and inductance are not uniform, and for this reason the current and voltage distribution is not truly sinusoidal.

If, now we double the applied frequency (or with the same applied frequency, increase the wire length to $\frac{\lambda}{2}$ approximately) another resonance point will be found (the natural wavelength of an unearthed wire), but this time a voltage antinode occurs at the base end of the wire, and the distribution of E and I is as shown in Fig. 143b. As before, the exact length of wire to give resonance is not equal to $\cdot 5\lambda$, but more nearly $\cdot 47\lambda$, the exact figure being largely dependent on the proximity of other objects and the method of supporting the free ends.

A further increase of length to three-quarters of a wavelength (or increase of frequency to $3f$) shows another resonance with maximum current at the base and a distribution of E and I as shown in Fig. 143c. In fact a whole series of resonance points will be obtained, the tuning at the odd quarter wavelengths producing maximum *current* at the base, and tuning at the half wavelengths maximum *voltage* at the base. Such an aerial operating at a multiple frequency is called a harmonic aerial, the exact length of aerial in terms of the harmonic frequency becoming more nearly the exact theoretical figure for the higher harmonics.

If the aerial had uniformly distributed inductance and capacity and no resistance or radiation, then its behaviour would be identical with that of the open-ended feeder discussed in Chapter VI, and hence we can get useful approximate ideas as to how the reactance of an aerial changes by applying the feeder analysis to our aerial problem.

It was shown that the reactance of an open-ended feeder was given by

$$X = -j\sqrt{\frac{L}{C}} \cot \omega h \sqrt{LC}$$

where h is the length of feeder, or, in this case, the height of the aerial. Maximum current at the foot of the aerial will evidently occur when this is zero, that is when

$$\omega h\sqrt{LC} = \frac{\pi}{2} \text{ or } \left(n + \frac{1}{2}\right)\pi,$$

n being an integer.

In the case of a straight wire in clear surroundings, $\dfrac{1}{\sqrt{LC}} = c$, the velocity of light;

$$c = \frac{\omega}{2\pi} \cdot \lambda$$

Hence $$\omega h \frac{2\pi}{\omega\lambda} = \frac{\pi}{2} \text{ or } h = \frac{\lambda}{4}$$

as we have already seen to be the case with the assumptions made. Professor Howe has shown that

$$L = 2\left(\log_e \frac{2h}{d} - 1\right) 10^{-9} \text{ henries,}$$

is a close approximation for the inductance per cm. of a vertical wire near earth and

$$C = \frac{1}{\left(2 \log_e \dfrac{2h}{d} - 1\right)} \times \frac{1}{9 \times 10^{11}} \text{ farads}$$

for the capacity per cm., where d is the diameter of wire in cms., and h the length, also in cms.

It is of interest to observe that the characteristic impedance of such an aerial will be

$$Z_o = \sqrt{\frac{L}{C}} = 60\left(\log_e \frac{2h}{d} - 1\right).$$

For instance, an aerial 15 metres long, 0·25 cm. diameter, will have a characteristic impedance of

$$Z_o = 60\left(\log_e \frac{3{,}000}{\cdot 25} - 1\right) = 510 \text{ ohms.}$$

If we have a wire of definite length and work out its constants as above, then its frequency reactance curve will be as shown by the reactance curves in Fig. 144, this curve showing the multiple tuning points just discussed. Thus at every odd quarter wavelength its reactance measured between the base end and earth (base reactance) is zero, and (since the aerial has been

assumed resistanceless) the base resistance is zero. For every even harmonic the base resistance is infinity and hence any circuit used to feed the aerial between the base and earth must be capable of impedance variation between wide limits, if it is to match the aerial.

Since, of course, an aerial always has losses, the frequency/impedance curve will not comprise the discontinuous reactance

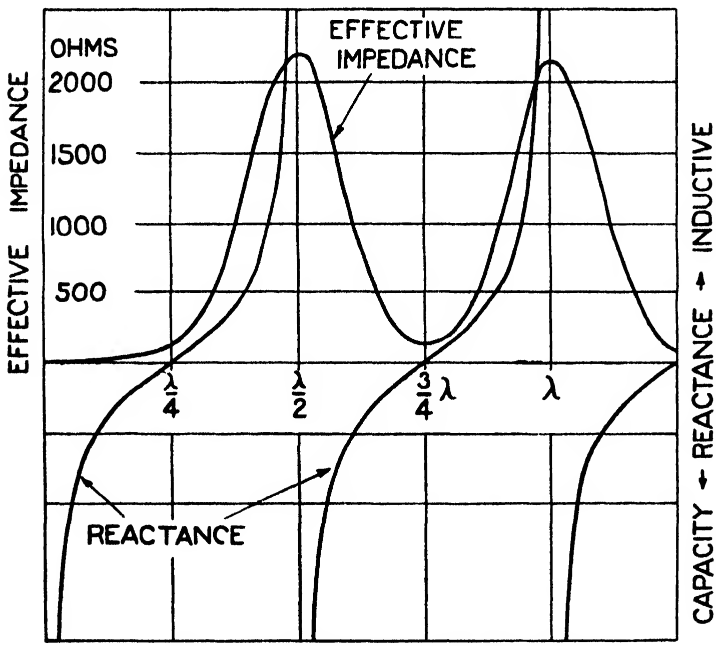


FIG. 144. Aerial Reactance and Impedance Curves.

curves shown in Fig. 144, but a modified curve as shown by the impedance will be obtained, whose shape depends upon the aerial's characteristic impedance and its losses as will be shown later. For the time, however, let us consider what the reactance changes imply.

For instance, if an aerial is at exactly quarter wave resonance (or odd multiples of this) with the applied frequency, then its base reactance is zero, and a suitable coupling would be a series tuned circuit of L_1, C_1, which will also have zero reactance, as shown in Fig. 145a. If the aerial is somewhere near quarter

wave resonance, but not exactly in tune, then the base circuit should also be de-tuned until the sum of the reactances is zero, that is, with the assumptions made :

$$\sqrt{\frac{L}{C}} \cot \omega h \sqrt{LC} = \omega L_1 \quad \frac{1}{\omega C_1}$$

This equation can be solved by plotting the reactance curves of the aerial and of the tuning circuit, but the result would be very approximate, due to the assumptions made regarding the aerial reactance, and in practice the correct adjustment of an aerial is easily found by varying the tuning circuit until maximum current is obtained in the earth connection.

At the $\frac{1}{2}\lambda$, λ, $\frac{3}{2}\lambda$, tuning points, the aerial acts as a parallel resonant circuit, its base reactance is zero, but base resistance infinity and the only circuit to which we can couple it efficiently is, therefore, one of parallel inductance and capacity, whose resistance at resonance is also infinite, the arrangement being shown in Fig. 145b.

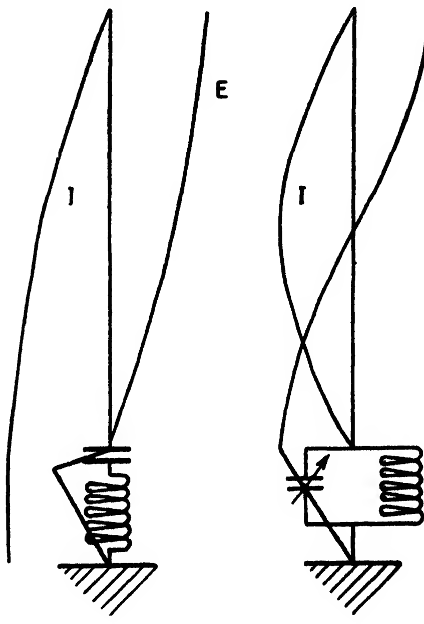


FIG. 145. Tuning $\lambda/4$ and $\lambda/2$ Aerials.

In the same way as before, if the aerial is nearly, but not quite a half wave (or multiple thereof), the parallel circuit needs to be de-tuned. In this case no convenient aerial current reading is possible, except with medium and high-power circuits or with cage aerials, since at the base end the feed current is only very small, and correct adjustment is best found for transmitting work by observing the loading on the transmitter when the aerial is coupled.

We will now consider how resistance modifies the above cases. The aerial is not a loss-free feeder, and if a quarter wavelength or more has a considerable radiation resistance as well as losses, which vary with its mode of oscillation, its proximity to earth, and the type of earth.

The aerial disposes of energy in the following ways:

(1) Radiation of electromagnetic waves.

(2) Losses in the earth and due to eddy currents in neighbouring conductors.

(3) Losses in aerial wire and in insulators, etc.

Since these losses are proportional to the square of the current, they can be replaced in our consideration of the aerial

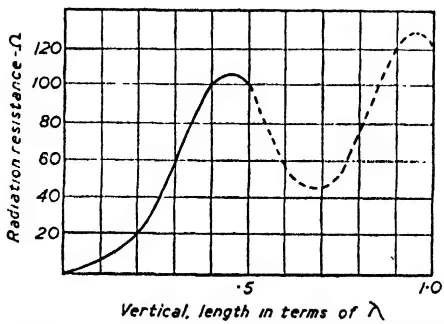


FIG. 146. Variation of Radiation Resistance with Length of Aerial.

as a circuit by an equivalent, entirely-fictitious resistance located at any convenient point in the aerial. Since, however, the aerial is carrying a stationary wave, the value obtained for this resistance will vary, being least at current antinodes and increasing to high values near the current nodes.

To avoid ambiguity, therefore, it is necessary to specify the point at which this effective resistance is placed, the current antinode usually being chosen.

Actual values for the total resistance of aerials referred to the point of maximum current are difficult to calculate, and not easy to measure even at medium frequencies. We can derive the radiation component, which varies as shown in Fig. 146, from which we note that a $\frac{\lambda}{4}$ aerial on a perfectly conducting earth has a resistance of 36·6 ohms, whereas the radiation

resistance for a $\frac{\lambda}{2}$ aerial is 104 ohms. It is clear from the curve why good earth systems are so important with aerials small compared with the wavelength, and less so with aerials commensurate with $\frac{\lambda}{2}$. In the case of an aerial of $\frac{\lambda}{4}$ the total resistance might very well be as high as 50 ohms, and since the effective resistance is taken at the base of the aerial, the base resistance is also 50 ohms.

In the half-wave case the feed is at the base of the aerial, where the voltage is a maximum, whilst the effective resistance R is referred to the centre of the aerial where the current is a maximum. Hence the base resistance will not be the same as R but some higher value.

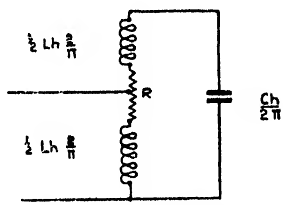


FIG. 147. Equivalent Circuit of Aerial.

We can get at the base resistance of a half-wave aerial approximately from the following considerations. Let us try to construct a closed resonant circuit composed of *concentrated* inductance, capacity and resistance which shall behave in the same way as the half-wave aerial having *distributed* inductance and capacity. Note that the aerial and the closed circuit will only behave in the same way for the exact frequency which produces half-wave resonance of the aerial. Note also that the concentrated inductance and capacity we are going to use in our closed circuit represent values which could not actually be measured in the aerial, the only measurable quantity being the base resistance which we are trying to calculate.

Suppose that the closed circuit carries in all its parts a current I equal to the maximum current in the aerial—that is, the current at the centre. If the aerial (of height h cms.) has a uniformly distributed inductance of L per cm, then its total effective inductance would be Lh, if it carried the current I throughout. To allow for the actual distribution we should use the mean value of the current which is $I \cdot \frac{2}{\pi}$ (for a half-sine

wave) and hence the total effective inductance L_e is $Lh. \frac{2}{\pi}$ and this is the value of the inductance in our equivalent closed circuit, as shown in Fig. 147.

Dealing with the capacity of the aerial in a similar way, we shall need in this case to consider the two quarter-waves separately, because the voltages are of opposite sign. If C is the capacity per unit length, $\frac{Ch}{2}$ would be the effective capacity if the voltage was uniform but $\frac{Ch}{2} \cdot \frac{2}{\pi}$ if the voltage is distributed in a quarter-sine wave. The two quarter-waves are in series and therefore the total effective capacity C_e is $\frac{Ch}{2\pi}$.

The half-wave aerial is at earth potential at its centre point and, therefore, the base impedance is really measured between one end and the centre of the aerial, and the arrangement of the equivalent circuit is as shown. It is seen to be a parallel-resonant circuit tapped half-way down, and its equivalent resistance at resonance is therefore

$$R_e = \frac{\omega^2}{R}\left(\frac{L_e}{2}\right)^2 = \frac{\omega^2 L_e^{\,2}}{4R}.$$

Since $\omega^2 L_e C_e = 1$, this may be written $\frac{L_e}{4\,R\,C}$, or in terms of L and C,

$$R_e = \frac{Lh\frac{2}{\pi}}{4R\,\frac{Ch}{2\pi}} = \frac{L}{R\,C}.$$

If we take as an example a half-wave aerial 1,500 cm high and of 0·25 cm diameter, then, from the formulæ on p. 259,

$$\frac{L}{C} = 4\left(\log_e \frac{3{,}000}{\cdot 25} - 1\right)^2 \times 9 \times 10^2 = 258{,}000$$

and if R is 120Ω then the base resistance, $\frac{L}{CR} = 2{,}150\Omega$, shown at $\lambda/2$ in the impedance curve of Fig. 144.

It will be seen that an increase in the diameter of wire used

will decrease L and increase C. It is reasonable to assume R to be unaffected by the change because R is mainly radiation resistance (on short waves) and there will be negligible change in conductor and earth loss.

In consequence, the base resistance decreases if the wire diameter is increased, but the effect is not very marked unless a cage aerial or tubular conductor is employed. For instance, doubling the diameter in the example given would only reduce R_e to 1,730 ohms, but for a cage of 200 cm diameter R_e would become 280 ohms.

The effect of resistance will, of course, influence the shape of the current-voltage waves along the aerial, indicated in Fig. 145. Although the stationary wave current is chiefly dependent upon the aerial's equivalent resistance at the maximum current point, the feed current is inversely proportional to the impedance at the feed point, which in the cases shown in Fig. 145 will be between the aerial base and earth. Although all aerials carrying a $\lambda/2$ stationary wave will have a total resistance at the current loop of somewhat more than 100 ohms the base resistance at resonance, and the sharpness of tuning, will depend considerably upon aerial design. Thus an aerial consisting of a long thin wire will have a very high base-impedance. In such a case the feed current will be very small in comparison and the tuning, of course, very sharp. With a cage aerial, however, since the base impedance is now much lower, the feed will be quite appreciable and will modify the current waveshape.

This feature is useful when we have to design an aerial to work over a band of frequencies, as in television. For instance the Alexandra Palace aerials consist of triangular cages of 38 cm sides, and for television reception it is common practice to use a tubular half-wave aerial which is good both from an electrical and mechanical point of view.

Thus, in the case of an aerial having resistive loading due to radiation and losses, the value of impedance varies over extremely wide limits as the frequency is changed (see Fig. 144), and it can be seen that to tune an aerial to wavelengths between the natural resonance points necessitates not merely the alteration of reactance of tuning circuit, but an essential alteration of the type of coupling circuit.

In the case where one wavelength only is required, we do not need to have a variable tuning circuit. An alternative method of tuning the aerial can be arranged, therefore, namely through a length of wire (called a tail) and a coil, as shown in Fig. 148. For if the curve of Fig. 144 be referred to, it is clear that an aerial longer than $\frac{\lambda}{2}$ but less than $\frac{3}{4}\lambda$ has a capacitive reactance between its base and earth, and thus between these limits an inductive loading will be required to tune to resonance.

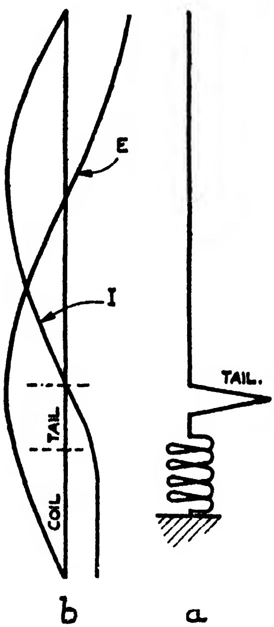


FIG. 148. Tuning an Aerial by a Coil.

The conditions very near the half-wave point are so influenced by high base-resistance, that it is found very difficult to tune an aerial only just a fraction longer than $\frac{\lambda}{2}$ by inductive loading, and it is customary to add a tail between the aerial proper and the loading coil, as shown in Fig. 148a; curve 148b shows the stationary wave on aerial tail and coil, the current in the latter being uniform. This tail should be carried horizontal to earth and preferably doubled back on itself so as to reduce radiation therefrom, and its length should not be less than $\frac{\lambda}{10}$ or more than $\frac{\lambda}{8}$. If the tail is too short it is difficult to tune as previously explained, whereas if it is too long, the base current is higher, more radiation from the tail is experienced, and earth losses are greater.

Radiation from an Aerial

The radiation from the simplest radiator—the dipole—has already been discussed in Chapter IV.

Actual aerials are not dipoles because their dimensions are comparable with the wavelength radiated and the current in the various parts is not uniform. In the long-wave case,

where the aerial is usually only a fraction of the wavelength, the dipole equations may be used in obtaining field strengths at a distance by inserting an equivalent height to allow for

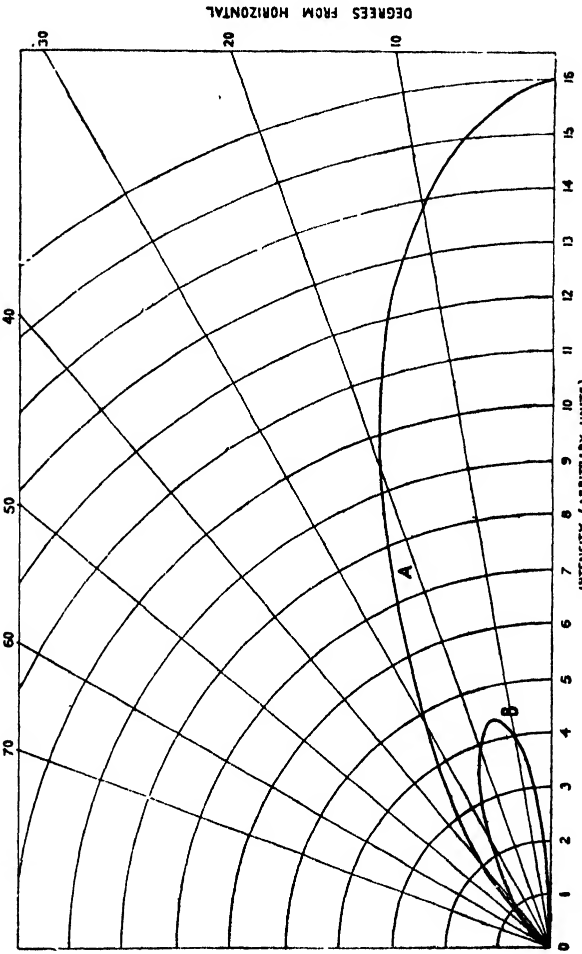


FIG. 149. Polar Diagram of $\lambda/2$ Vertical Aerial, Just Above the Earth.

the current distribution. Such a method is not applicable to the short-wave aerial, however, since it may be more than a wavelength high, will have a much less uniform current-distribution and will not have a polar diagram of cosine form. It will be necessary to treat the actual aerial as built up of a number

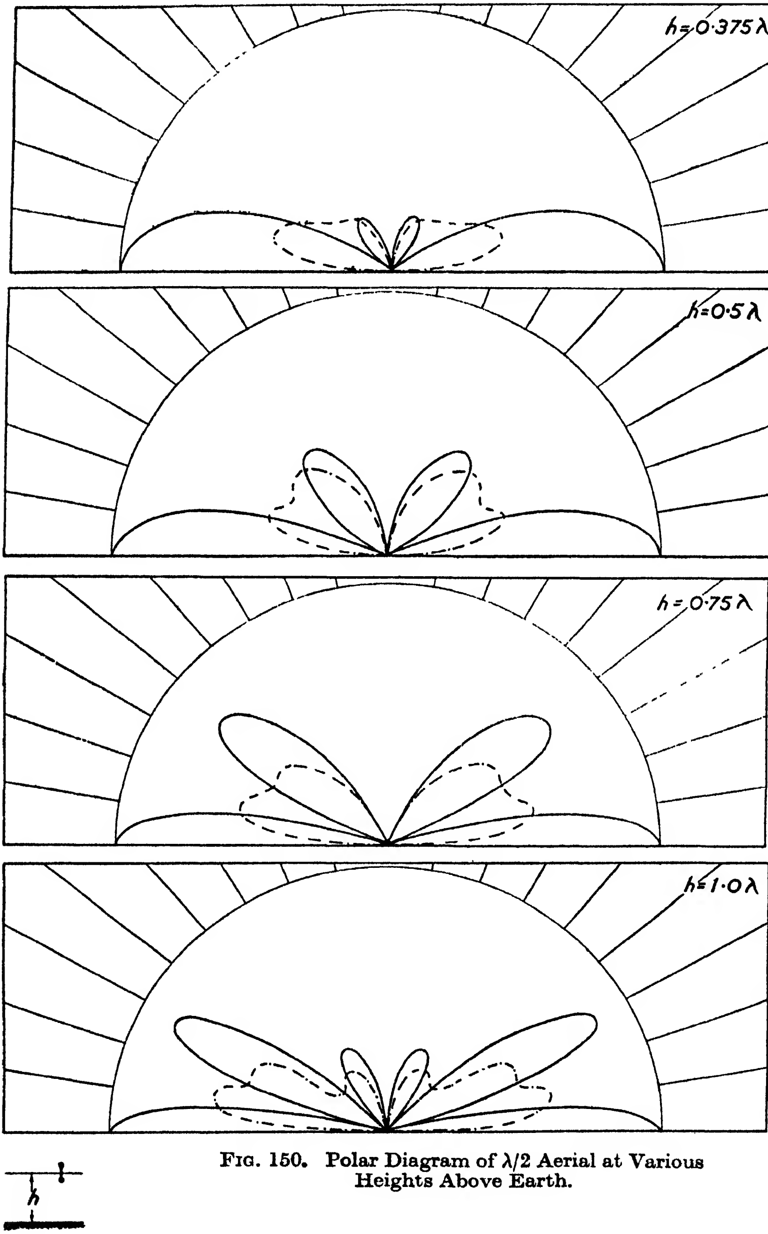


Fig. 150. Polar Diagram of λ/2 Aerial at Various Heights Above Earth.

of elementary lengths each having a cosine polar diagram and carrying currents appropriate to their position in the aerial.

The effect of the earth will be allowed for by an image aerial, also split up into elements presumed to be carrying currents of magnitude and phase dependent upon earth constants, in the way discussed in Chapter IV.

If a $\frac{\lambda}{2}$ aerial is assumed to be in free space, then the radiation resistance can be shown to be 73Ω. This value is frequently used for ultra-short wave aerials elevated to a number of wavelengths above earth.

For example, the zenithal polar diagram for a half-wave aerial just above a perfectly conducting earth is shown in Curve A, Fig. 149, and is seen to be much sharper than that of a dipole.

From this polar diagram a value for the radiation resistance has been deduced by Ballantine using the method explained on page 267, and a value of 104 ohms obtained. By the same method the radiation resistance of a quarter-wave aerial with its lower end earthed can be worked out and is found to be 36·6 ohms. Diagrams for a half-wave aerial at various heights above a perfectly conducting earth are shown by the full curves in Fig. 150.

These results assume a sinusoidal distribution of current which is everywhere in the same time phase—that is, a perfect stationary wave, whereas, since energy is being radiated, the actual current must be a combination of a stationary wave and a travelling, energy-conveying wave. In the case of aerials made up of thin wires, the base impedance is high and the travelling wave is therefore very small compared with the stationary wave and the assumed sinusoidal current distribution is a good approximation. For cage and other large diameter aerials, however, the travelling wave is a larger proportion of the total current and the assumption is not so justified but is usually made on account of the simplification produced.

Effect of Imperfectly Conducting Earth on Zenithal Polar Diagram of Aerial

If the finite conductivity and dielectric constant of the earth is taken into account, the polar diagram for a given aerial is

dependent upon frequency. The polar diagram, since it is dependent on the "image" theory which assumes optical reflection at the earth's surface, does not take account of the "surface wave." In other words, the diagram is really only approximate for great distances and ignores energy transmitted along the earth's surface, this being quickly absorbed in the case of short waves.

The polar diagram for a half-wave vertical aerial radiating a 22-metre wave and situated just above the surface of earth having likely values of conductivity and dielectric constant, has been calculated and is shown by curve *B* in Fig. 149, where it is compared with the polar diagram when the earth is perfectly conducting. It will be seen that whereas the diagram, when the earth is considered to be perfectly conducting, has its maximum value in a horizontal direction, when likely earth constants are assumed there is no horizontal radiation at the wavelength considered.

The radiation resistance of an aerial producing a polar diagram such as curve *B* can be found by comparing the area of curve *B* with the area of curve *A*, since the radiation resistance in the latter case is known to be 104 ohms, and the polar diagrams are the same in all horizontal directions. Note that a direct comparison of polar diagram areas does not give a correct result because the field strength at a low angle on the surface of the hemisphere is the same over a zone having considerable area whilst the field strength at a high angle only exists over a zone of small radius. When this difference of area has been allowed for the method shows the radiation resistance to be 31 ohms whilst the total effective resistance of such an aerial has been measured experimentally and found to be about 165 ohms so that its efficiency for long distance communication is about 18%.

Much of the energy which appears to have been lost is probably radiated in the surface wave and theory and experiment agree that raising the aerial greatly improves the efficiency.

It has already been explained that approximate polar diagrams of vertical short wave aerials are best obtained by considering the image aerial to carry an equal current in anti-phase for the lower angles of elevation and an equal current in-phase for the higher angles.

Curves for half-wave aerials are shown in Fig. 150, in which the full curves are for an in-phase image and dotted curves for an anti-phase image.

To assess the merits of aerials, particularly those above 0·5λ in height, we must consider both the radiation resistance in comparison with the total resistance and also the polar diagram. This is most important, since radiation resistance alone will give an imperfect picture of what is happening. Thus for a one wavelength aerial, R_r is seen, from Fig. 146, to be 120Ω, but the radiation in the horizontal plane from such an aerial is zero. We have therefore dotted the curve above 0·5λ.

Zenithal Diagram of Harmonic Aerials

We can get increased height merely by making the wire longer and having a number of half-waves on it forming a

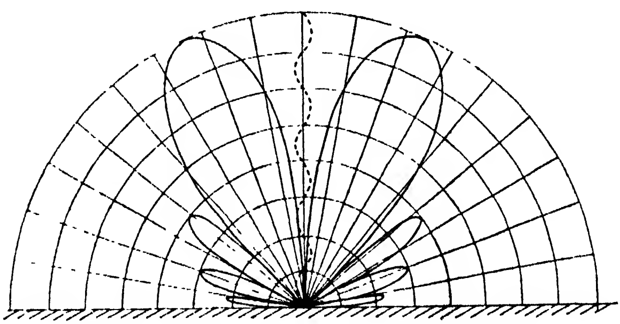


FIG. 151. Polar Diagram of 4/λ Vertical Aerial.

harmonic aerial, but if we do this we get increased high angle radiation, whilst the radiation from the various half-waves will cancel out at low angles, because adjacent half-waves are in phase opposition. The zenithal polar diagram of a four wavelength, vertical harmonic aerial is shown in Fig. 151. But it has already been pointed out (in Chapter V) that it is of considerable advantage when transmitting to concentrate the energy radiated into a sharp " beam " at a low angle because we have only a definite amount of energy available and must use it in the most effective way. When receiving it is equally important to use an aerial to receive over only a small sector having a small angle to earth, to avoid multiple effects. Thus

to get increased height effectively and yet obtain concentration of field at low zenithal angles necessitates a special type of aerial.

Tiered or Stacked Aerials

To gain this object C. S. Franklin developed a tiered aerial, earlier types suppressing the alternate half-wave radiation, and later types using the alternate half-wave portions to assist in producing radiation in the required direction.

Franklin's first tiered aerial comprised a series of half-wave aerials with coupling condensers between each, the small capacity between the ends of the wires being sufficient to transfer energy at the very short waves used and reverse the phase. Then followed the phasing-coil type in various forms whose shape was designed " to concentrate alternate half-wavelength portions of the wire within a small space so that there is practically no radiation from these portions," to quote from the Patent Specification. In practice the proportions of wire lengths are not half-waves, but empirical rules have been found to give the best results.

Although such types of aerial marked a great advance in the efficiency of short wave aerials, they were not ideal as the abrupt change of characteristic from open aerial to " lumped " inductance-capacity caused reflections at the various points, and resulted in a reduction of radiation from the top portion of the system, this being undesirable. To overcome these defects, a so-called " uniform " aerial was produced.

Franklin Uniform Aerial

In order to get the greatest possible concentration of radiation at a low zenithal angle with a given total height of aerial, the ideal arrangement is a vertical wire carrying a uniform current in the same phase. An aerial approximating to the ideal can be made in a variety of ways, one type being indicated in Fig. 152, which shows that each successive half-wave wire is folded back on itself in such a manner that the radiation from its central part assists radiation from adjacent wires. The radiation from the tips of the phasing wire cancels the radiation from the tips of the adjacent wires, but since it is the sections carrying the maximum current which are chiefly productive of radiation

the elimination of radiation from the ends does not matter. By such means we attain almost to the ideal of a uniform current aerial as indicated in the figure, and thus utilise the available height in the most economic manner possible.

The increase in the length of wire carrying useful current results in the power radiated (and of course the power drawn from the transmitter) being greater than in the case of a half-wave aerial.

If we have high masts and only a given amount of power available, it may be more advantageous to concentrate this in a few aerials well away from earth rather than in more aerials, some of which are close to earth. This can be done by folding between the feeder proper and aerial the lower length of wire, so that it becomes nearly non-radiating, a type used by Franklin being shown in Fig. 152.

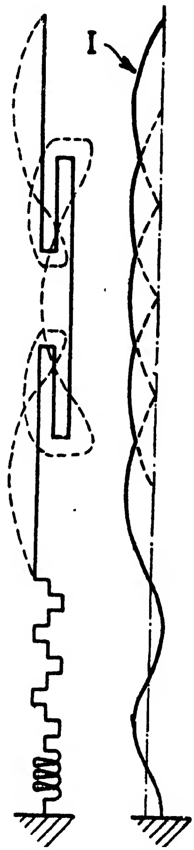


Fig. 152. Franklin Uniform Aerial.

Vertical Polar Diagrams of Tiered Aerials

We can determine approximately the vertical polar diagrams of a system of tiered aerials by the following method, which assumes the aerial to carry a uniform current of the same phase throughout its length. The effect of the earth is taken into account by an image aerial of opposite phase, since we are mainly interested in the low angle radiation.

Consider a vertical aerial of length $l = n\lambda$ (Fig. 153a) with a corresponding image aerial as shown. If we imagine the aerial to be divided into elementary lengths of aerials all carrying a uniform current, then the field produced at a point P, distant from O, at any angle $\theta°$ from the ground, will be determined by the vector sum of the individual fields. If the field due to one element be E, and there are x elements, we have to sum up x vectors, each of which is slightly out of phase with the

preceding one, because the distance between each element and P is not the same. These vectors form approximately the arc of a circle $A'B'$, as shown in Fig. 153b, the chord being the resultant, the shape of this arc depending upon the angle being considered and the aerial length.

Now since the arc is the arithmetic sum of the fields its length is constant (say unity) and hence the ratio $\frac{\text{chord}}{\text{arc}}$ determines the relative field at any given direction. Consider Fig. 153b. The phase angle between the first and last vector equals ϕ radians, the angle at the centre of the circle subtended by the arc. The ratio

$$\frac{\text{chord}}{\text{arc}} = \frac{\sin \frac{\phi}{2}}{\frac{\phi}{2}}.$$

Hence, if we find the phase difference between the first and last elements of the aerial in terms of the zenithal angle θ we can obtain the relative field strengths from the above expression.

The extra distance to be covered by the wave from the lowest element is OC (Fig. 153a). If this distance is λ, the phase difference would be 2π radians, hence the actual phase difference ϕ is given by

$$\frac{2\pi}{\lambda} . OC$$

$$\text{or } \phi = \frac{2\pi}{\lambda} n\lambda \sin \theta$$

$$= 2\pi n \sin \theta.$$

So far we have considered only the aerial, but we must now allow for the effect of the image. As explained previously, we can take into account the effect of the earth by assuming an image aerial carrying a current in phase opposition to that in the actual aerial when we wish to obtain an approximate diagram of low angle radiation and one carrying a current in phase for high angle radiation. The in-phase case merely requires that the length of the aerial be taken as twice its actual length.

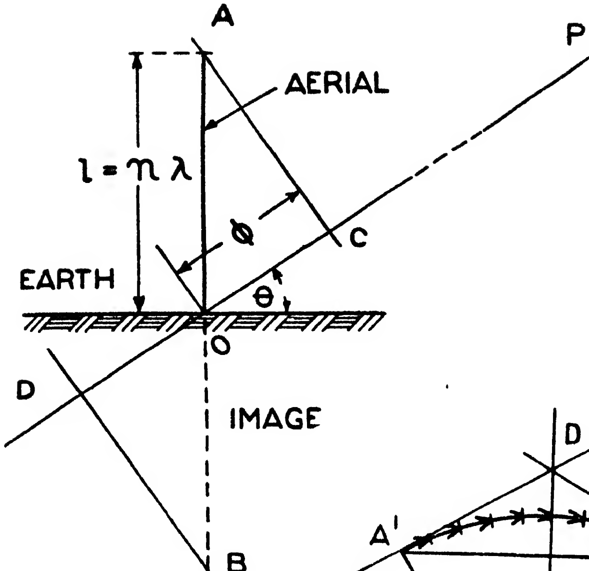


Fig. 153a.

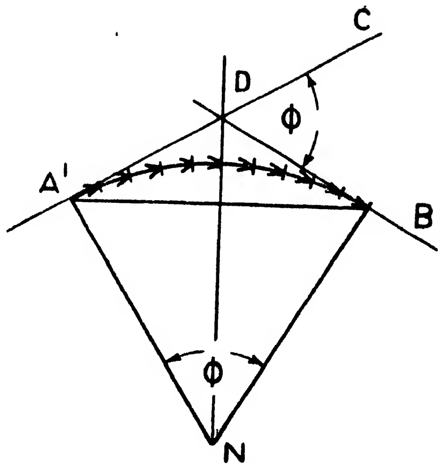


Fig. 153b.

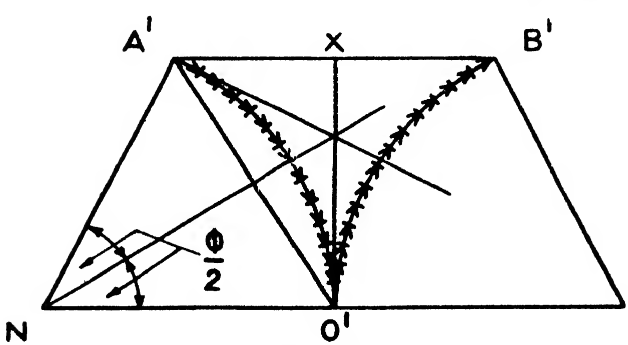


Fig. 153c.

Fig. 153. Calculation of Polar Diagram of Aerial.

If the image oscillates in opposite phase, adjacent vectors at O (the earth) are exactly in anti-phase and a similar but reverse arc of vectors for the image is obtained as shown in Fig. 153c, the resultant vector being $A'B'$.

Draw $O'X$ tangential to $O'A'$ to cut the line $A'B'$. It will bisect $A'B'$ at right angles and $O'X$ will also be at right angles to $O'N$ since $O'N$ is the radius of the circle and $O'X$ the tangent.

Since angle $NO'X$ is a right angle and the angle $NO'A$ is equivalent to $\pi/2 - \phi/2$, the angle $A'O'X$ equals $\frac{\phi}{2}$. Now $A'X = A'O' \sin \phi/2$, therefore $A'B' = 2A'O' \sin \frac{\phi}{2}$. But $A'O'$ is the chord and equals $\dfrac{\sin \frac{\phi}{2}}{\frac{\phi}{2}}$ (since the arc is assumed of unit length)

Thus
$$A'B' = \frac{2 \sin \frac{\phi}{2}}{\frac{\phi}{2}} \cdot \sin \frac{\phi}{2}$$
$$= \frac{2 \sin^2 \frac{\phi}{2}}{\frac{\phi}{2}}.$$

The maximum possible field strength at P is, however, twice that due to the aerial alone, owing to the image. Also we have not, so far, allowed for the fact that the radiation from each element of the aerial will follow a cosine law in the vertical plane. Allowing for these facts we have

$$\frac{\text{Actual field strength at } P}{\text{Max. possible field strength}} = \frac{\sin^2 \frac{\phi}{2}}{\frac{\phi}{2}} \cos \theta$$

where the relationship between ϕ and θ has been shown to be $\phi = 2\pi\, n \sin \theta$.

As it is usually more convenient to work in degrees than radians, the above formula should be modified thus :

$$\frac{\text{Actual field strength at } P}{\text{Max. possible field strength}} = \left\{ \frac{\sin^2 \frac{\phi}{2}}{\frac{\phi}{2}} \cdot \cos \theta \right\} \times 57{\cdot}3$$

where ϕ and θ are measured in degrees.

We have already noticed that an aerial cannot carry a pure stationary wave, since there is energy radiated along its length. When several half-wave aerials are in series, as in the tiered aerial, then there must be a considerable travelling wave entering the aerial at its foot in order to supply energy to all the sections, so that at the foot of the aerial the travelling wave is a considerable component of the resultant, whilst near the top it is negligible.

In consequence of this the maximum radiation from a tiered aerial is not perpendicular to its length, even if earth effects are neglected, and the actual tilt of the main lobe of the polar diagram is dependent upon the frequency supplied to the aerial.

The effect is specially marked in the Franklin uniform aerial where, by altering the applied frequency (or, alternatively, changing somewhat the lengths of the various portions of the aerial), the main radiation can be varied in direction considerably. Some experimental results indicated that a 6% wavelength change could swing the angle nearly 10°.

Connecting R.F. Lines to Vertical Aerials

In many cases there will be a considerable length of R.F. line between transmitter or receiver, and aerial, and the method of connecting aerial to line needs consideration.

The line should, of course, be terminated by a load representing a resistance equal to Z_0. In arrangements such as those shown in Figs. 145, 148, 152, where a loading coil can be employed, the impedance between any turn of the coil and earth will be approximately resistive, the value increasing as we tap up the coil. If the line is of the concentric type with the outer earthed, then the correct termination can evidently be obtained simply by tapping the inner conductor to the correct point on the coil.

Matching of an elevated $\frac{\lambda}{2}$ vertical aerial fed at the centre (which forms a load balanced to earth) to a concentric line (which is clearly unbalanced to earth) can be carried out by the arrangement shown in Fig. 161, where U.S.W's are used.

One method for matching an unbalanced aerial to a parallel-wire line is shown in Fig. 154. The length of wire AB, one end of which is connected to A, the base of the aerial, is $\frac{\lambda}{2}$ long, so that it reflects an equal reactance at the point B, but in opposite phase, and thus the feeder may be regarded as looking into a balanced, reactive impedance. The stub is adjusted to correctly terminate the line, as explained in Chapter VI. There will, of course, be stationary waves on the $\frac{\lambda}{2}$ length and it is folded in the manner shown in order to reduce radiation from it.

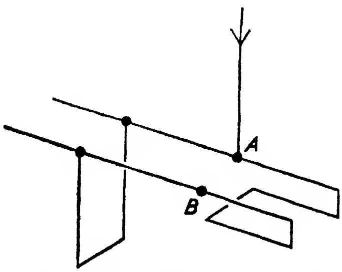


FIG. 154. Matching an Unbalanced Aerial to a Parallel-Wire Line.

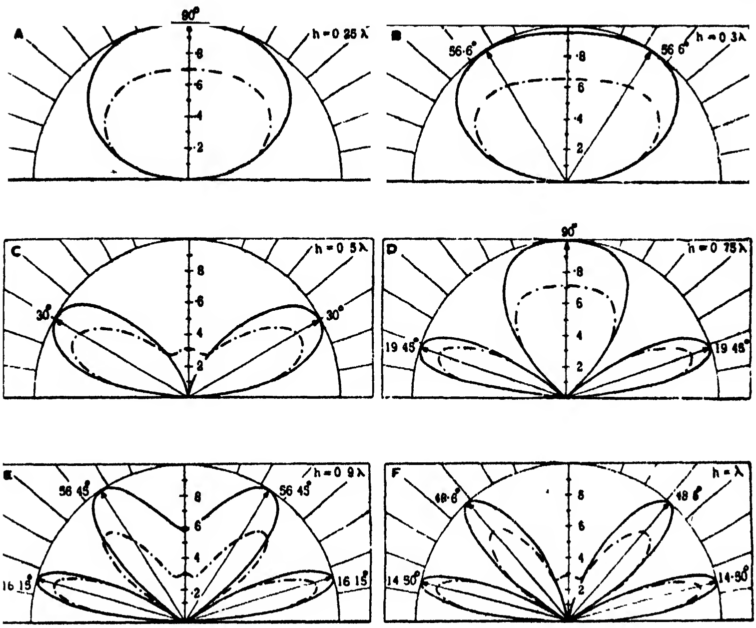
Horizontal Aerials

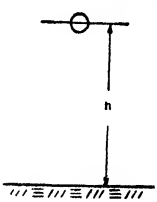
So far we have considered only vertical aerials, but horizontal aerials are also greatly used. They are usually fed through an unearthed feeder of the parallel-wire type, being either designed for high efficiency at a spot wavelength, or for moderate efficiency over a band of wavelengths, and a variety of simple arrangements is possible.

If the effect of the earth is neglected altogether, the polar diagram of a horizontal dipole is the same as for a vertical one, but turned through a right angle so that the zenithal polar diagram becomes the horizontal diagram and vice versa. The polar diagram in a horizontal plane through the dipole is therefore a "figure of eight" and the dipole is seen to be directional.

We have already noted that the electric and magnetic fields radiated from a dipole are perpendicular to each other and both

perpendicular to the direction of propagation. It follows that a vertical dipole radiates vertically polarised waves in all horizontal directions but the radiation in directions making





——— Perfect Conductivity $\sigma = \infty$

—·—·— Average Earth $\begin{cases} \sigma = 10^7 \text{ E.S.U.} \\ \kappa = 5 \text{ E.S.U.} \end{cases}$

Curves are deduced assuming mathematical dipole at height h above the earth as indicated.

A	$h = \cdot 25\lambda$	B	$h = \cdot 30\lambda$
C	$h = \cdot 50\lambda$	D	$h = \cdot 75\lambda$
E	$h = \cdot 90\lambda$	F	$h = 1\cdot 0\lambda$

FIG. 155. Polar Diagrams of Horizontal Aerials.

an angle with the horizontal contains both vertically and horizontally polarised components.

If the dipole is now turned through a right angle it will be seen that the horizontal radiation in a direction normal to the axis of the dipole will be horizontally polarised but that the horizontal radiation in other directions will contain both horizontal and vertical components.

Let us now consider the effect of placing a horizontal dipole

a short distance above a perfectly-conducting earth. The propagation of horizontally-polarised waves along the surface of such an earth would not be possible because it would involve differences of potential over a perfect conductor since the electric field is horizontal. Eddy currents would be produced, the field from which would exactly cancel out the original field. For usual values of the earth's conductivity the horizontal radiation is very small and the zenithal polar diagram for a given aerial position relative to earth is less dependent upon the earth's constants than in the case of a vertical aerial.

The polar diagram at distances large compared with the height above the earth may be found by the image method previously discussed for vertical aerials, but in the case of the horizontal aerial we must reverse the sense of image current as against that used for a vertical aerial over a similar earth, as has been explained in Chapter IV.

Fig. 155 shows a series of zenithal polar curves calculated for a "mathematical dipole" at different heights above a perfect and imperfect earth whose values of σ and κ are as indicated in the figure. From these curves it is very evident that as one is chiefly interested in low-angle radiation there is considerable advantage in raising an aerial a half wavelength above earth and that very little additional advantage will be gained by raising it above ; unless the aerial can be raised to a height of at least one wavelength or more. Aerials raised between a half and one wavelength above earth have, as can be seen, strong radiating properties at high zenithal angles. An interesting feature is that neither the earth conductivity nor the wavelength have a vast influence on the shape of the polar curves. Although these curves are calculated for an idealised dipole they are substantially true for the ordinary half-wave type of horizontal aerial.

The radiation pattern from a horizontal aerial in the earth's plane is roughly a figure of eight with maxima in line with the aerial, but with some radiation in a plane normal to the aerial instead of zero radiation. Such an aerial therefore is not so suitable for omni-directional working as is the vertical aerial.

It is known that whether waves are radiated from horizontal or vertical aerials, they are usually circularly-polarised when they leave the ionosphere, hence, at distances beyond the skip

area, either type of aerial may be used for reception from a transmitting aerial of either type. One type of fading is produced by changing polarisation and can be partially overcome by summing up the E.M.F.'s received on two aerials, one vertical and one horizontal, for tests show that interference fading on a horizontal aerial is often opposite to that on a vertical aerial near it, and receiving from the same station. Generally a horizontal aerial, orientated the correct way of course, provides a better signal/noise ratio than a vertical because certain electrical apparatus, such as ignition systems or sparking commutators on machines, produce highly damped E.M. waves of high frequency, which appear to be vertically polarised.

Feeding of Horizontal Aerials

We have already mentioned that the horizontal aerial has not very good omni-directional characteristics. Furthermore, when it is employed as a correct termination to a line (its most efficient condition) its tuning will be critical and it is therefore a " spot-wave " radiator. Since it is a matter of convenience to employ aerials which have a wavelength coverage, there are a number of ways in which this can be accomplished, but at a sacrifice of efficiency. We propose to discuss some of the more important types that are in use.

Thus Fig. 156a shows a horizontal half-wave aerial whose actual length l, as explained on page 258, will be less than $\frac{1}{2}\lambda$, usually $\cdot 47\lambda$ to $\cdot 475\lambda$, fed from a twin feeder connected to points XY on the aerial such that a correct feeder termination is obtained. The correct distance XY depends on the characteristic resistance of the feeder used and on the height of the aerial above earth. If a 600 ohm feeder is employed and the distance of the aerial above earth is $\frac{1}{2}\lambda$ or more, $XY = \cdot 125\lambda$, whereas if the aerial is nearer earth, say $\frac{1}{4}\lambda$, $XY = \cdot 1\lambda$ for the same feeder. It should be observed that the depth of the bight between the aerial and the commencement of the parallel-wire feeder proper has an influence on the tapping position XY and a length of bight $\cdot 15\lambda$ will be found correct for the dimensions previously given.

In the case of a transmitting aerial the feeder wires are run parallel (two 14-gauge wires separated 15 cm giving 600 ohms

is a common dimension), but with feeders for receiving work it is most desirable to use a transposed feeder, either by transposing the individual wires about every 2 metres, or by using a twisted flexible wire. An equally important precaution is to make sure that the wires run are exactly the same length, both these measures being desirable in order to prevent unwanted "pick-up" on the feeder.

Such an aerial is highly efficient at wavelengths near the natural tune of the horizontal wire, but the efficiency falls away rapidly on either side of the tune. Since the feeder is correctly terminated its length is not important.

An alternative "general purpose" arrangement, usually only

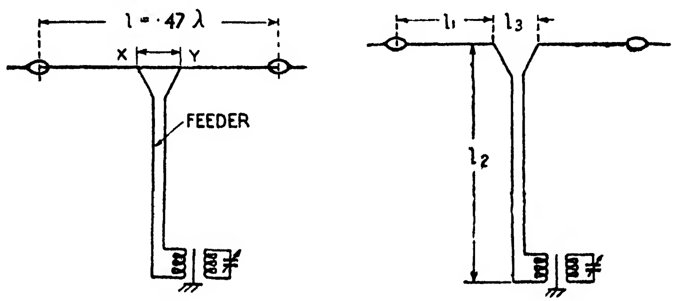


FIG. 156. Feeding Horizontal Aerials.

used for receivers, is shown in Fig. 156b. In this case the overall length of horizontal aerial, that is $2l_1 + l_3$, will be made approximately equal to one-half of the most important wavelength within the band it is desired to cover, and the dimensions of such an arrangement will be as follows: $2\,l_1 + l_3 = \frac{1}{2}\lambda$; the connecting piece inserted at $l_3 = \cdot 3\lambda$; and the length of feeder l_2 must not be longer than necessary, since it is for the most part incorrectly terminated, but it should be dimensioned such that it is $\cdot 5\lambda$ or $1{\cdot}0\lambda$ if tapped into a parallel resonant circuit at the base; or $\cdot 25\lambda$, or $\cdot 75\lambda$ if tapped into a series circuit at the base.

Such a "general purpose" aerial may be regarded as operating as arrangement (*a*) on and near the optimum wavelength, but at longer waves the feeder is not terminated, but now assists in the reception, and in consequence the length

of the feeder will determine the performance of the aerial as a whole. Thus we can regard the aerial as made up of two parallel aerials of length l_1 and l_2, whose various natural resonance wavelengths can be found as previously indicated, and knowing these we can determine the type of circuit required to match the aerial. Because there are two down leads, vertically-polarised interference will be eliminated, and in consequence such an aerial will give a somewhat better signal/noise ratio than a simple vertical wire.

In cases where a simple directive aerial is desired, it is possible to use the arrangement of Fig. 156b, but increase the length of each limb of the aerial to $\frac{1}{2}\lambda$. In this case the end of feeder will not be fanned out but run parallel right up to the aerial connections which are each now end fed (see below). Since the feeder line is now not correctly terminated even at the optimum wavelength of the aerial system, the length of feeder must be designed correctly to fit the type of terminating circuit at the base, as indicated on page 282.

The Zeppelin Aerial

A form of feeder-aerial combination popular with amateurs but not much used by commercial organisations is as shown in

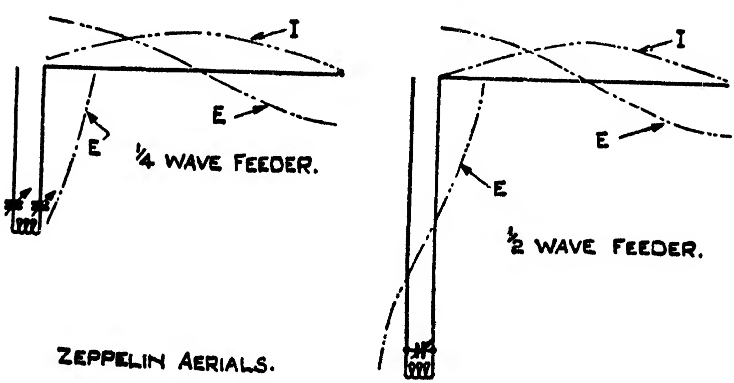


FIG. 157. "Zeppelin" Aerials.

Fig. 157, termed the Zeppelin aerial. It will be seen to consist of an aerial one half wavelength long (or a multiple thereof) and a parallel-wire feeder, one limb of which is connected to one end of the aerial.

Since the aerial is fed at a voltage antinode, a feeder must be provided of the correct length to produce the voltage antinode at its far end ; it is clear that a feeder of any length may be used, although it is highly desirable to keep it as short as possible consistent with the height of aerial, provided the ground or feed end be terminated by the correct tuning system to produce stationary waves as required.

If the curve "*O*," Fig. 92, page 179, be referred to, we can determine the type of circuit necessary at the feed end. For instance if a feeder just less than one quarter wavelength is used, the feed end will be capacitative and of small value, and a small coil coupling will suffice. Actually, series tuning condensers can be provided so that this can be adjusted to give maximum current in the feeder at this end. If a feeder nearly one half-wavelength long be used, since the feed end is now of high impedance, a parallel feed circuit will be required, the current in the feeder now being very small for optimum adjustments.

Thus in setting up a feeder and aerial of this type, having a knowledge of the length of the feeder in terms of the wavelength, we can decide easily upon the type of feed circuit to be used. It should be observed that a feeder system operated under such conditions will tend to radiate, but the radiation can be minimised by balancing the current in the two wires to exact equality. This point has previously been mentioned.

Single Wire Feeder

It is possible to feed a half-wave aerial through a single wire feeder and so adjust the position of the tap that the current in the feed wire is free from stationary waves. Consider a half-wave aerial as shown in Fig. 158, on which is shown stationary waves of current and voltage. Consider the impedance between any point *A* on the aerial and earth. We can regard two circuits as connected to the point *A*. An open circuited wire *AB* (rather less than quarter wavelength in the case shown) and a second open circuited wire *AC* (the same amount greater than one quarter wavelength). If now the reactance curves on page 260 are consulted, it will be observed that wherever the point *A* be chosen the reactance of the length *AB* is always equal and opposite to that of the length *AC*, referred to the

point A. If the tap is near the centre, these reactances are small and as the point A is moved away from the centre, the reactances rise to large values, but at all points the effective reactance is zero, and thus we are always " looking into " a resistive circuit wherever the point A be placed. The values of this effective resistance are low at the centre and rise to the high values associated with the base resistance of a half-wave aerial at the end.

In feeding an aerial through a single-wire feeder, provided the feeder is tapped at such a point along the aerial that it is " looking into " a resistance equal to its own surge resistance, the feeder termination is correct and no stationary waves are present in the feed wire.

As shown in a previous section a vertical wire of normal dimensions near earth has a characteristic impedance of about 500 ohms and thus the point " A " will be chosen to match this. A point about $\cdot 037\lambda$ along the aerial either side of the centre will be found to be nearly correct.

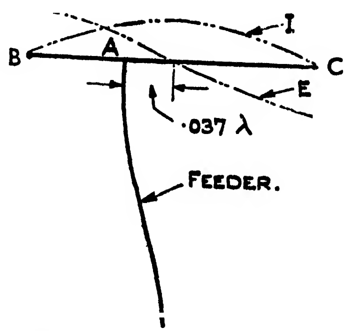


FIG. 158. Single-Wire Feeder to Aerial.

It is evident that there will be radiation from the single-wire feeder, but if the tapping point has been correctly adjusted to eliminate stationary waves, the current in the feeder will be much less than in the centre of the aerial and hence the greater part of the radiation takes place from the aerial, as desired. The adjustment of the tapping point to eliminate stationary waves is very critical and it will be realised that maladjustment is more serious than with the Zeppelin aerial because of the radiation from the feeder that results.

The Quadrant Aerial

Where omnidirectional working and a frequency spread are required, a fairly simple form of horizontal aerial, known as the Quadrant Aerial, can be used. This takes the form of a right-angled horizontal Vee, as shown in Fig. 159, fed through

a feeder at the apex of the Vee by a parallel-wire feeder. Where a spot frequency only is required, a single-wire Vee is used and the feeder matched exactly, but when a wave coverage is required, as is more usual, the arms of the Vee are made of cage formation, and the feeder employed should have a charac-

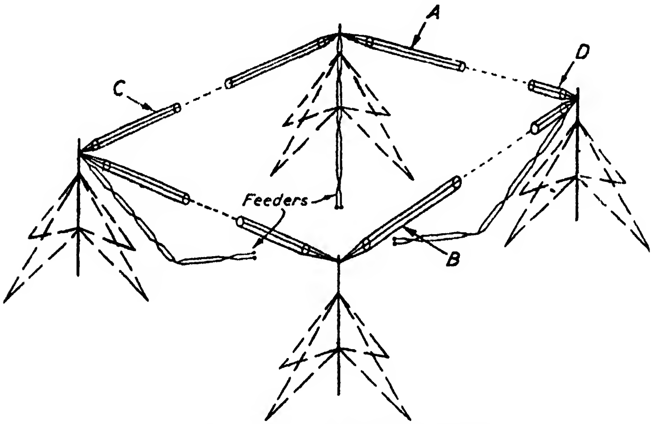


FIG. 159. The Quadrant Aerial.

teristic impedance between 400 and 600 ohms. This can take the form of a twin, open feeder, or, for reception only, a concentric line is often used with an appropriate transformer between the aerial terminals and the feeder. With a cage type aerial and feeder of value as stated, it is possible to obtain a 2/1 wave coverage fairly efficiently. The length of each side of the Vee will be made ·45λ of the mid-band frequency, and the height of the aerial should be based on the average requirements of the service. Since height governs the angle of incidence of the main radiation, as is seen from Fig. 155, its correct determination is important. The polar diagram of radiation in the horizontal plane is roughly the superposition of two figures-of-eight at right angles to each other and in space quadrature for the mid-band frequency, the diagram

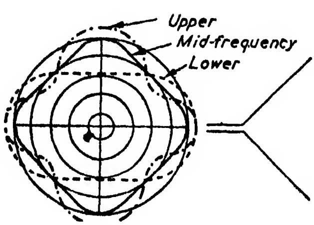


FIG. 160. Polar Diagram of Quadrant Aerial.

departing from this at other frequencies, Fig. 160 showing the mid-frequency diagram, and those at the frequency limits, from which we can observe that in all cases the aerial is fairly omnidirectional.

Fig. 159 shows a battery of four quadrant aerials, having cages 1 metre square, designed to cover the whole short wave band. Thus *A* will cover 2·0 to 4·0 Mc/s ; *B*, 3·5 to 7·0 Mc/s ; *C*, 6·0 to 12·0 Mc/s ; and *D*, 10 to 20 Mc/s. The side of each quadrant will be proportioned as indicated above and the height based upon average requirements. Actually, 100-feet masts are found to be very suitable.

Some Points in Design and Use of Aerials

Enough has been said to show that aerials should be designed for the wavelength at which it is desired to work, but this is, however, not always possible and efficiency may have to be sacrificed to practical convenience. A marine transmitter or receiver, for example, will have to work over a large range of wavelengths and we cannot rig up a special aerial for every wave and may even have to work on the ship's long-wave aerial. For reception, the aerial may be coupled to the receiver through a length of low-capacity cable, the aerial circuit being untuned. The type of horizontal aerial described in a previous section is also suitable.

For transmission, the ship's main aerial may be used as a harmonic aerial, coupling its lower end through a series or parallel circuit as described in the first section of the chapter. If we know the length of aerial available, we can very quickly see at what wavelengths such an aerial will be efficient, and possibly by a small alteration of length, which will not effect its working on long waves, efficiency can be obtained at the particular short waves desired.

It will be appreciated that long "leading in" wires, or "trunks," which are a common feature of the long wave system, may be most inefficient on short waves, as the capacity of these leading in arrangements may be sufficient to by-pass most of the current. Their length may be favourable to the formation of a large stationary wave due to reflection at the deck insulator, where the inductance and capacity per unit length change greatly in value. Where possible it is always

better to use a feeder, commencing the aerial proper where the surroundings are clear, although the use of such will necessitate a terminating circuit at the bottom of the aerial. If this cannot be done, correct termination at the transmitter end of the trunk may enable more power to be fed to the aerial.

Aerials for Ultra-short Waves

We have so far discussed aerials mainly from the short-wave point of view, but the principles apply, of course, to ultra-short waves and most of the types discussed are equally applicable to these wavelengths. Various modifications are possible, however, when the wavelength is very short.

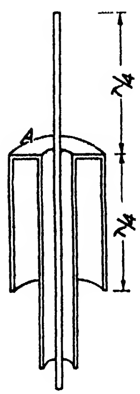


FIG. 161. U.S.W. $\lambda/2$ Aerial Fed from Concentric Line.

Since it is so easy and profitable to raise the aerial several wavelengths above earth, it is usual to use elevated $\frac{\lambda}{2}$ aerials, rather than arrangements such as the Uniform Aerial. Such aerials, whether vertical or horizontal, will be centre-fed. If a concentric line is used, this may be adapted to the balanced load formed by the $\frac{\lambda}{2}$ aerial, in the way shown in section in Fig. 161. A copper cylinder, of considerably greater diameter than the outer of the line, acts as the lower half of the aerial, whilst an extension of the inner acts as the upper half. The impedance placed across the line at A is about 75Ω and is therefore a suitable termination for the line.

It will frequently be the case that a high modulation frequency has to be dealt with (in television, for example) and hence on wavelengths of a few metres the aerial needs to work efficiently over a large percentage bandwidth. The impedance at the drive point should be fairly uniform over the band, otherwise the feeder cannot be correctly terminated over the band of frequencies involved and reflections will occur, which are particularly undesirable in television.

We have already seen that the driving-point impedance and Q of an aerial is lower if its diameter is increased and tube aerials are frequently employed. A number of special shapes are used, all designed to lower L and increase C, and examples

are shown in Fig. 162. Such aerials may be made up of continuous surfaces (as Fig. 162c) or be composed of a cage of

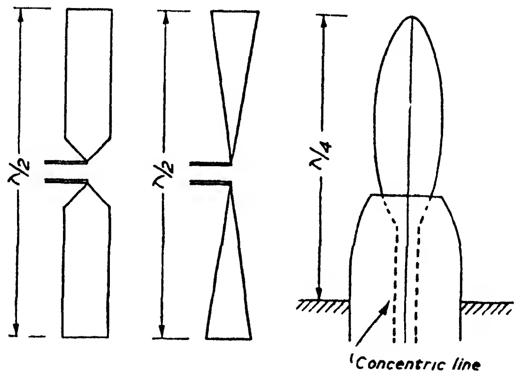


FIG. 162. Wide-band U.S.W. Aerials.

wires. A folded wire aerial, as shown in Fig. 163, is sometimes employed, the two wires close together having much the same effect on the impedance as an increase in diameter. Methods for calculating the driving-point resistance and reactance have been developed by Moullin and by Schelkunoff, amongst others.

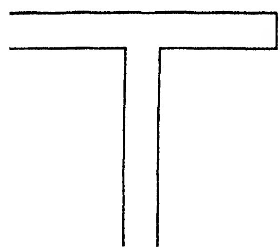


FIG. 163. Folded-Dipole Aerial.

On centimetric waves, $\frac{\lambda}{2}$ or similar aerials are seldom used without some kind of reflector to form a beam and such arrays are discussed in the next chapter.

Slot Aerials

Suppose that we cut a slot, approximately $\frac{\lambda}{2}$ long, in a very large sheet of metal and feed current into the sheet by a parallel-wire line, as shown in Fig. 164.

It will be seen that the two halves of the slot form $\frac{\lambda}{4}$

short-circuited lines and currents will flow in the sheet. Thus there will be a p.d. across the slot and an electric field concentrated across the slot, which will be a maximum at the centre and zero at the ends. This field will also fringe out on both sides of the sheet.

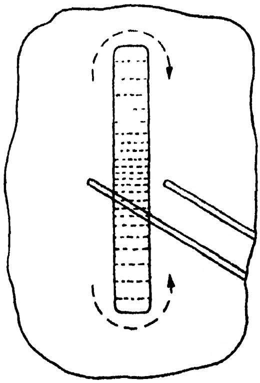


FIG. 164. Slot Aerial.

A comparison of the field distribution with that which would have been produced if the piece cut out to form the $\frac{\lambda}{2}$ slot had been used as a $\frac{\lambda}{2}$ aerial, will show that the magnetic and electric fields are interchanged but are otherwise of the same form. The slot therefore acts as an aerial but the radiation is polarised at right angles to that produced by a $\frac{\lambda}{2}$ aerial occupying the same position.

The impedance across the centre of the $\frac{\lambda}{2}$ slot can be shown to be a resistance of 485Ω. The wider the slot the less the falling off in current circulating round the slot if the frequency is shifted from $\frac{\lambda}{2}$ resonance. Hence, widening the slot has the same effect upon bandwidth as increasing the diameter of a $\frac{\lambda}{2}$ aerial.

Normally we wish the slot to radiate on only one side of the sheet and this can be arranged by boxing up one side of the slot, a hemisphere about $\frac{\lambda}{4}$ in radius being suitable. The impedance is then doubled.

This equivalence between slot and wire radiators is not, of course, limited to the $\frac{\lambda}{2}$ resonant slot, but any arrangement of wire radiators has its counterpart in slots, though some of the arrangements are practicable and useful, whereas others are not.

Slot aerials have found considerable application on centimetric waves and may be useful in very high-speed aircraft as slots cut in the fuselage and filled with solid dielectric would not offer any wind resistance.

In our example, the slot was excited by a line, but slots may be cut in the sides of wave guides and will function in the same way, provided that they are in the correct direction for the type of wave being propagated down the guide.

If a slot is cut so that it is across the original current flow on the face of the guide, then it can be seen that the new current distribution will approximate to that shown in Fig. 164 and therefore the slot will radiate in the same manner. The properties of slots in wave guides have been studied by Watson

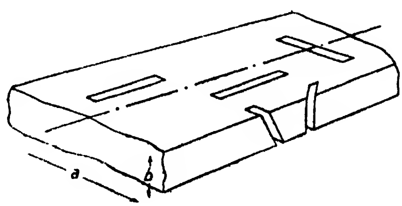


FIG. 165. Slot Radiators in a Wave Guide.

and a number of other workers, using $H_{1,0}$ waves in rectangular guides.

The slots can be cut in the broad face (*a*, see Fig. 165) in the longitudinal direction and the radiation from these will increase from zero if they are along the centre-line (where there is no transverse current) to a maximum at the edge of the face. The phase of the radiation is opposite on the opposite sides of the centre-line.

A transverse slot can also be cut with its centre in the middle of the broad face, as there is a longitudinal current there. Inclined slots can also be used.

Approximately transverse slots can also be cut in the narrow face (*b*) of the guide. If exactly transverse, they would not radiate because there is no longitudinal current in this face, but inclined slots are useful. The *b* wall of the guide will not normally be $\frac{\lambda}{2}$ long but the slot can be cut as shown in Fig. 165.

CHAPTER IX

AERIAL ARRAYS

THE term "aerial array" has been given to combinations of aerials arranged to produce some special directional effect. The whole array may in some cases be used simultaneously for the transmission (or reception) of more than one signal.

The advantages of directional transmission and reception have, of course, been realised from the early days of wireless, but efficient systems have only become possible since short waves came into use, as it is only possible to produce any efficient transmitting or receiving directional aerial-system when the size of the array is large compared with the wave-length being used.

General Requirements

When a long distance point-to-point service is under consideration, the horizontal polar diagram at the transmitter should be narrow, and the ratio of forward to back radiation should be as great as possible to concentrate the available energy into a narrow beam. The vertical diagram also should be sharp and it may be desirable to be able to change the directive properties in the zenithal plane so as to be able to select the particular ray predominant at any one time.

At both transmitter and receiver there is a limit to the narrowness of the horizontal diagram, because we have evidence that the apparent direction of the transmitting station does not always remain quite constant.

Since the arrays are probably the most expensive part of a short-wave station, it is very desirable in the interests of economy to be able to couple a number of receivers to the same array, or to be able to use a transmitting array on more than one frequency. This means that the polar diagram should not be dependent upon precise tuning of any part of the array.

The direction of maximum horizontal radiation should not be too rigidly dependent upon the direction in space of the

array. For it is evidently an advantage if, by a simple electrical adjustment, an array can be biassed to radiate in a different direction.

Some special cases may, of course, have peculiar requirements. An interesting example is the type of array required for the telephone service from Rugby to Atlantic liners. In this case the horizontal diagrams must evidently be wide enough to cover the area of ocean over which it is desired to maintain communication, and in the case of the longest wave array, which is used for the shorter distances, the beam has a large spread and is arranged to radiate principally at fairly high angles, as it is the high angle radiation which returns from the ionosphere in the shortest distance.

The most obvious way of forming a beam is to use a parabolic reflector with a single aerial at the focus.

This method of producing a beam, which was the first employed, is now used only for ultra-short waves, as newer systems have been developed which possess many advantages, chief of which is the simpler mechanical construction, and the greater gain of field strength produced by division of energy among a number of aerials.

The modern aerial array system, of which C. S. Franklin was the originator, obtains its " beam " effect by the grouping of a number of radiating elements fed by current in such phase as to produce an interference pattern in space giving the required directive properties.

In general, array systems are of two main types, the " Broadside " type and the " In-line " or " End-Fire " type. In both cases the array consists of a long line of radiating elements, but whereas the Broadside type concentrates the energy normal to the line of radiators, the later is most active along the aerial line.

Before we consider the various systems in use, it will be of interest to illustrate the interference pattern produced by two spaced omnidirectional aerials.

Polar Diagram of Two Spaced Aerials

Consider the field produced at P (Fig. 166) by two aerials A and B (assumed to be point sources of radiation) spaced $n\lambda$ metres apart, where λ is the wavelength in space corresponding

to the frequency of the currents in the aerials, it being assumed that P is so far away that lines joining A and B to P are parallel. Evidently, if the currents in the aerials are in the same phase, the fields due to each aerial will not be in phase at P, because of the difference in distance, from P to each aerial.

This difference in distance is most conveniently expressed by the phase angle it produces between the fields at P due to

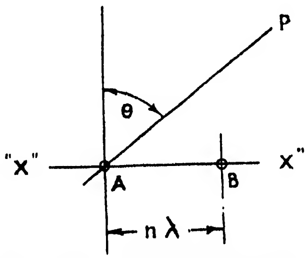


FIG. 166. Two Spaced Aerials.

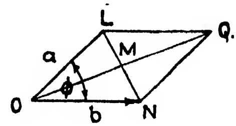


FIG. 167. Vector Diagram for Resultant Field.

the two aerials, and this phase angle ϕ will be termed the "space-phase."

It will be seen that the difference between the lengths PA and $PB = n\lambda \sin \theta$, and, since a difference in length of λ would cause a phase angle of 360°, the phase angle actually produced will be $\frac{360°}{\lambda} n\lambda \sin \theta$, or $\phi = 360° \, n \sin \theta$.

If the currents in A and B are in phase and produce fields a and b at P which are equal numerically, then the vector sum at P is OQ, as shown in Fig. 167.

Now $\qquad OQ = 2OM$ and

$$OM = ON \cos \frac{\phi}{2}.$$

Hence $\qquad OQ = 2b \cos \dfrac{\phi}{2}.$

Applying these formulæ to spacings $\frac{\lambda}{8}, \frac{\lambda}{4}, \frac{\lambda}{2}$ and λ, we get the polar diagrams shown in Fig. 168, top line.

In many cases, however, the two aerials are supplied with currents which are out of phase in time so that there is a

time-phase of α degrees. For such cases the resulting phase angle, ϕ' say between the two fields, will be dependent upon whether the time-phase α is additive to, or must be subtracted from, the space-phase ϕ. For directions to one side of the normal the two will add, whereas on the other side they will subtract, that is

$$\phi' = \alpha \pm 2\pi n \sin\theta \text{ radians}$$

or

$$\phi' = \alpha \pm 360 n \sin\theta \text{ degrees.}$$

If the current in the more distant aerial is lagging behind that in the nearer one, then evidently ϕ' is the sum of the

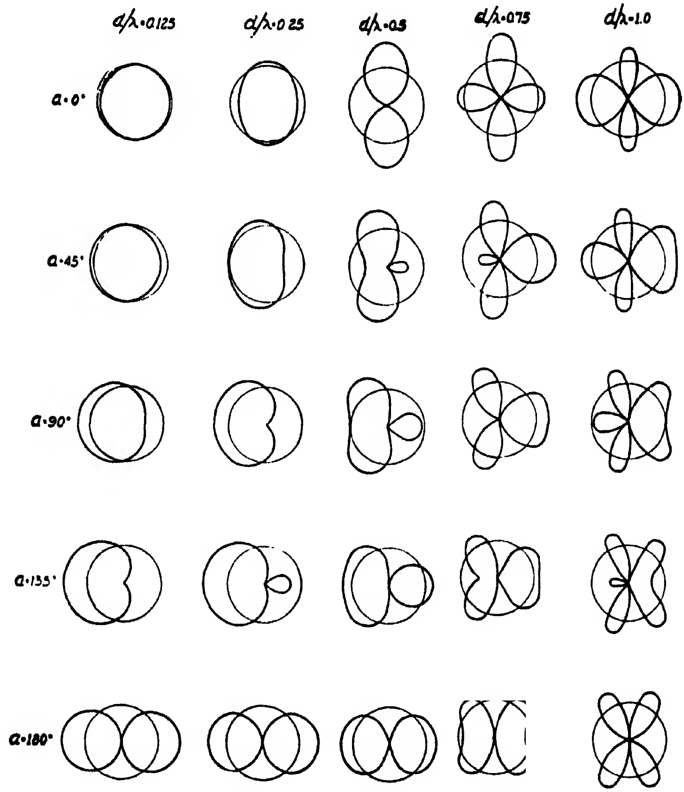


FIG. 168. Polar Diagrams—Two Spaced Aerials.

space-phase and the time-phase, whereas, if the current is leading, ϕ' is the difference.

With different values of spacing and time-phase, polar diagrams such as those shown in Fig. 168 are produced, the circles showing polar diagrams of a single aerial supplied with the same power. Time-phases between 180° and 360° would give similar diagrams but reversed in sense along the XX axis. From these curves it is observed that with zero and 180° of time-phase, diagrams bi-symmetrical both about the line of spacing XX and the line normal to it are produced. Further, with zero time-phase maxima are always produced in a direction normal to the line XX, whereas with 180° time-phase, minima are always produced normal to XX. For phases in between 0° and 180° (and 180° and 360°) diagrams assymmetrical to the normal line are produced, such diagrams having optimum uni-directional properties along XX when the time-phase is 90° or 270°.

The Inter-action of Two Adjacent Aerials

The polar diagrams obtained above are correct if the currents in the two aerials have actually the relative phases specified, but it is necessary to realise that in practice two such aerials would be coupled together, due to each being in the field of the other, and we must now consider this coupling effect.

Thus, if we first tune each aerial in the absence of the other and then apply to both of them E.M.Fs. in the same phase, each aerial will induce an E.M.F. in the other, the phase of which will depend upon the spacing. The currents flowing will be due to the resultant of applied and induced E.M.Fs.

The problem can be worked out as a coupled circuit, introducing the notion of a mutual impedance Z_{12} between aerials 1 and 2. The calculation of Z_{12} is a complex matter, involving, as it does, integrating the effect of the total field (and not merely the radiation component) produced by one aerial on each element of the other, but the values of Z_{12} for various aerials and spacings have been worked out. As an example, Fig. 169 (derived from the paper of G. H. Brown) shows the resistive and reactive components of the mutual impedance between two $\frac{\lambda}{2}$ aerials, both at the same height and sufficiently

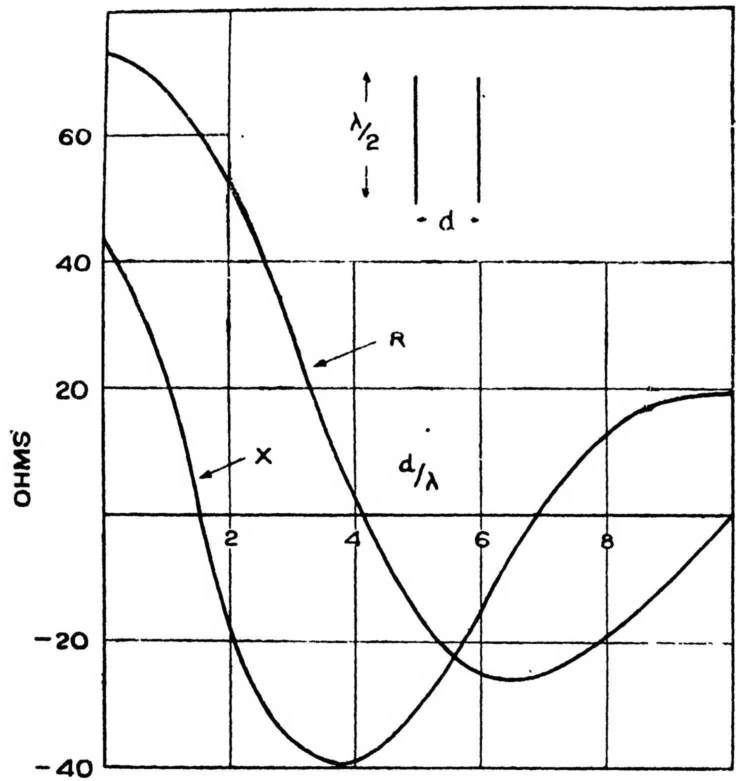


FIG. 169. Mutual Impedance between Two Aerials.

elevated that the effect of the earth on the mutual impedance may be neglected.

Similar curves for other cases are given by Brown and by Carter. Supposing Z_{12} to be known, then as in other coupled circuits :

$$\left.\begin{aligned} E_1 &= I_1Z_1+I_2Z_{12} \quad (1) \\ E_2 &= I_2Z_2+I_1Z_{12} \quad . \quad . \quad (2) \end{aligned}\right\} \text{These are vector equations.}$$

If $E_1 = E_2$ in magnitude and phase and both aerials are independently tuned, so that $Z_1=Z_2=R$, then $I_1=I_2$ in magnitude and phase. Hence the polar diagrams would have the same shape as found previously, but it is now seen that the aerials when together are no longer in tune, in the sense that

the impedance each presents to the circuit from which it draws its power is no longer a pure resistance. The currents are therefore different than for one aerial alone. The impedance is $\frac{E_1}{I_1}$ and from (1) is Z_1+Z_{12} (since $I_1=I_2$). Let $Z_{12} = R_{12}+jX_{12}$, then the aerials can be again brought into tune by altering their length so that Z_1 becomes $R-jX_{12}$. The aerials will now form a resistive load of value $R + R_{12}$ (R_{12} can be positive or negative, depending upon the spacing). Since the currents and effective resistances are not the same as for one aerial by itself, the power radiated is not twice that radiated by one aerial alone.

Aerial and "Parasite"

It is frequently convenient to make use of reflector or parasitic aerials which are not supplied directly with power, but the currents flowing in them are caused entirely by the E.M.F.s induced from neighbouring aerials. The relationships now become

$$E_1 = I_1Z_1 + I_2Z_{12} \quad \quad \quad (3)$$
$$0 = I_2Z_2+I_1Z_{12} \quad \quad \quad (4)$$

From which
$$\frac{E_1}{I_1} = Z_1 - \frac{Z^2_{12}}{Z_2}$$

$$I_1 = \frac{E_1}{Z_1 - \frac{Z_{12}}{Z_2}} \quad \quad \quad (5)$$

$$I_2 = -\frac{E_1 Z_{12}}{Z_1Z_2 - Z^2_{12}} \quad \quad \quad (6)$$

Since these are vector equations, they give us the magnitude and phase of the currents and we can then construct the polar diagram by the method previously given.

A very wide variety of effects are evidently possible, since we can vary Z_{12} (by altering spacing) and both Z_1 and Z_2 (by altering aerial tuning—adjusting the length, for example).

It is not easy to decide by inspection what kind of polar curve is likely to result from a given arrangement and a fairly tedious calculation is necessary.

The polar curve of an exactly-tuned $\frac{\lambda}{4}$ aerial and parasite spaced $\frac{\lambda}{4}$ apart has a shape between that for $\alpha = 90°$ and $\alpha = 135°$ in Fig. 168.

If aerial and reflector are both resonant and the spacing is varied we obtain the curves of Fig. 170 for field strength in forward and backward directions. It will be seen that the

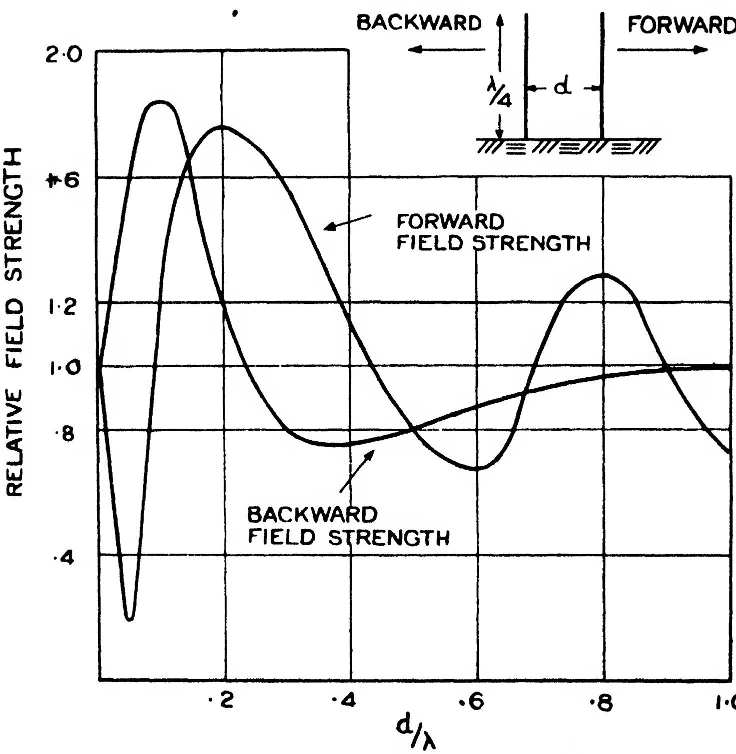


FIG. 170. Field Strength from Two Aerials.

parasitic aerial can act as a " director " as well as a " reflector " as the spacing is varied.

If the parasitic aerial is varied slightly in length, so that Z_2 has reactance as well as resistance, then larger differences between forward and backward radiation are possible.

It will be easier to follow the method of determining the polar curve if an example is taken. Suppose that the parasitic aerial is $0{\cdot}1\lambda$ from the driven aerial and that P (see Fig. 171) has been shortened until it has a capacitive reactance of 20Ω.

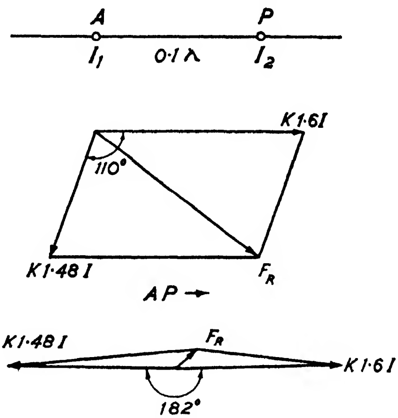


FIG. 171. Driven and Parasitic Aerial.

Since the change of length will be small, we shall use the curves of Fig. 169 (strictly applicable only to two $\lambda/2$ aerials) to find Z_{12}. This will be seen to be $67 + j22\Omega$.

We shall assume that the resistance of the aerials is all radiation resistance and that the small change of length of P has left its resistance at the $\frac{\lambda}{2}$ value of 73Ω. Hence $Z_1 = 73\Omega$ and $Z_2 = 73 - j20\Omega$.

Substitution in equations (5) and (6) gives

$$I_1 = 17{\cdot}9 \times 10^{-3}\, E_1 \underline{/59{\cdot}5^\circ} \text{ amperes}$$

$$I_2 = 16{\cdot}6 \times 10^{-3}\, E_1 \overline{\diagdown 86{\cdot}9^\circ} \text{ amperes.}$$

The power going into the array is given by

$$E_1 I_1 \cos\phi = E_1^2 \times 17{\cdot}9 \times 10^{-3} \cos 59{\cdot}5^\circ \text{ watts,}$$

and this is all radiated. If a single $\frac{\lambda}{2}$ aerial, carrying a current I, is to radiate the same power, then

$$E_1^2 \times 17{\cdot}9 \times 10^{-3} \cos 59{\cdot}5^\circ = I^2 \times 73$$

or $$I = 11{\cdot}2 \times 10^{-3}\, E_1 \text{ amperes.}$$

It follows that $I_1 = \frac{17 \cdot 9}{11 \cdot 2} I \underline{/59 \cdot 5^\circ} = 1 \cdot 60\, I \underline{/59 \cdot 5^\circ}$ amperes

and $\quad I_2 = \frac{16 \cdot 6}{11 \cdot 2} I \overline{\diagdown 86 \cdot 9^\circ} = 1 \cdot 48\, I \overline{\diagdown 86 \cdot 9^\circ}$ amperes.

The phase angle between the currents is 146°, I_2 lagging.

Along the line joining the aerials, the space-phase is 36°, and hence in the direction AP produced, the phase angle between the two components of the field will be 146° — 36° = 110°.

From the vector diagram we have

$$F^2{}_R = k\, I^2 \times 3 \cdot 14$$

and hence the gain in power compared with that from a single $\frac{\lambda}{2}$ aerial is 3·14 times, in this direction.

In the direction PA produced, the phase angle will be 146° + 36° = 182° and the gain is about 0·3 times.

It will be seen that the shortened parasitic aerial is a "director" and this is a general result, though the ratio of backward to forward radiation depends, of course, upon the spacing and length of the parasitic aerial.

If we had lengthened the parasitic aerial till it had an inductive reactance of 20Ω, we should have got a fairly similar value for the power gains but now the large radiation would be in the PA direction, so that the parasitic aerial would be acting as a "reflector."

When the aerial and parasitic aerial are used for reception (the aerial being connected to the receiver) the directional properties are similar to those for transmission but the actual performance is evidently dependent upon the way in which the receiving circuits load the aerial.

Polar Diagram of a Line of Radiators

The foregoing discussion relating to two single radiators illustrates the interference principle, but the diagrams obtained show equal energy concentration in various directions and are therefore not suitable for directional working. We can, however, eliminate the energy concentration in all except one main direction by employing a number of radiators spaced along the line.

For instance, consider a line of eight aerials spaced each one quarter wavelength apart as shown in Fig. 172, these eight aerials making an array line of length between the first and

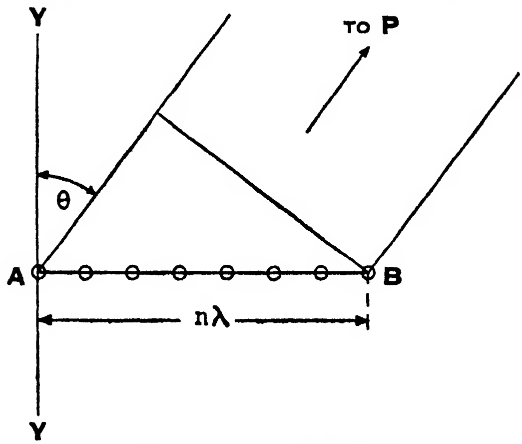


FIG. 172. Line of Radiators.

last aerial of $n\lambda = 2$ wavelengths. It will be found that such a line of closely-spaced aerials will concentrate the energy into

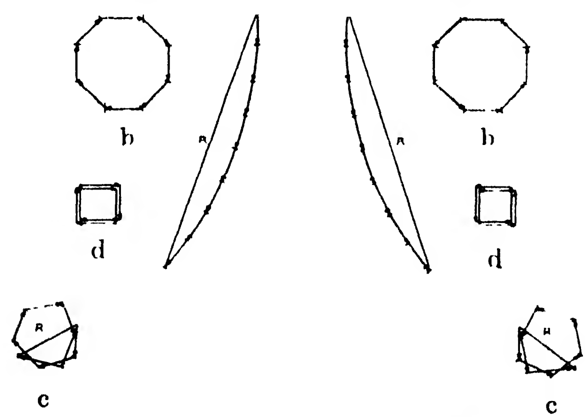


FIG. 173. Vectors for Line of 8 Aerials.

one main beam and a number of subsidiary " tails " each side of the line, and, moreover, the main beam can be made to assume any angle with the line of the aerials depending upon the time-phase of currents with which the aerials are supplied.

The field at a distant point, P, will, of course, be the vector sum of the fields due to each aerial. If we consider the important case when all the aerials carry currents in phase with each other, i.e. zero time-phase, then the maximum field as shown by the central vector (Fig. 173) will obviously be along a line YY, normal to the line of aerials. At directions other than normal, and making, say, any angle θ with it, to the right or to the left, the vector group will become wrapped up either in a clockwise or anti-clockwise direction as shown by Fig. 173a, b, c and d, right and left. Thus, for small angles of θ, a vector figure such as a will be produced. At greater deviations the vector sum produces a zero resultant such as shown in Fig. 173b (for the case of $n\lambda = 2$, θ will be 30°). Thus the field has diminished from unity at $\theta = 0°$ to zero when $\theta = 30°$.

Considering still greater deviations from the normal, the vectors become still more wrapped up to produce vector figures as shown in Fig. 173c, where the resultant is equal to the diameter of a circumscribed circle, and 173d where the vector diagram becomes wrapped up twice giving a zero resultant, this occurring for $\theta = 90°$.

General Formula for Polar Diagram of a Line of Radiators

As an alternative to drawing the vector diagram to scale, we can obtain a general formula. Suppose there are N aerials, all carrying equal currents in the same phase, with a spacing of d metres between adjacent aerials.

Then the phase angle ϕ between the components F_1, F_2 of the field at a distant point due to adjacent aerials will be

$$\phi = \frac{d \sin \theta}{\lambda} \cdot 360°.$$

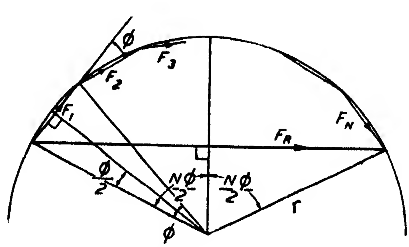


FIG. 174. Derivation of General Formula.

From Fig. 174 we see that the resultant field strength, F_R (the closing side of the polygon), is given by

$$F_R = 2r \sin \frac{N\phi}{2}$$

whilst F_1, F_2, etc., are each given by

$$F_1 \quad 2r \sin \frac{\phi}{2}.$$

Consequently
$$F_R = F_1 \cdot \frac{\sin \frac{N\phi}{2}}{\sin \frac{\phi}{2}}$$

When $\theta = 0°$, the field strength will be NF, and if this is taken as the reference value the field strength in any direction θ is evidently

$$F_R = F_o \frac{\sin \frac{N\phi}{2}}{N \sin \frac{\phi}{2}} \qquad (7)$$

Values of θ which make $\phi = 180°$, $360°$ etc., will be directions in which $F_R = 0$.

The numerator will evidently pass through maxima when $\frac{N\phi}{2} = 90°$, $270°$, etc., and, if N is, say, 4 or more, then the denominator will be changing so slowly compared with the numerator that these values of $\frac{N\phi}{2}$ will give the direction of the secondary lobes with fair accuracy.

If we maintain the length of the array line the same but alter the number of aerials (and, therefore, the spacing), the vector diagrams are but little altered in shape. We can, therefore, suppose our actual aerials replaced by a very large number placed very close together, when the vector diagrams will resolve into arcs and circles instead of polygons. This artifice, originally adopted by E. Green, has already been discussed on page 274 in connection with the zenithal polar diagram of a vertical aerial, and it will be evident that in the case of the array $n\lambda$ long, the field strength in a direction

making an angle $\theta°$ with the normal to the array line is related to the maximum field strength along the normal by

$$\frac{\text{Field at } P}{\text{Max. field}} = \frac{\sin\frac{\phi}{2}}{\frac{\phi}{2}} \quad . \quad . \quad . \quad (8)$$

where $\phi = 2\pi n \sin\theta$ (radians)

or $360\, n \sin\theta$ (degrees)

Minima will occur where ϕ, the phase angle between the field from the first and last aerial, is 2π or multiples thereof, namely at such angles that

$$\sin\theta = \frac{1}{n}, \quad \frac{2}{n}, \quad \frac{3}{n}, \quad \frac{4}{n}, \text{ etc.}$$

and maxima will occur at approximately angles such that

$$\sin\theta = 0, \frac{3}{2n}, \quad \frac{5}{2n}, \quad \frac{7}{2n}, \text{ etc.}$$

and the amplitude of the side loops relative to that of the main loop (unity amplitude) will be of value

$$\frac{2}{3\pi}, \quad \frac{2}{5\pi}, \quad \frac{2}{7\pi}, \text{ etc.}$$

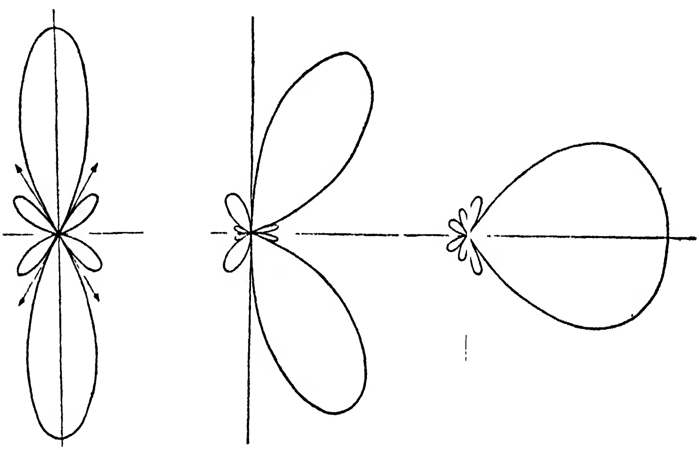
The polar curve for a two wavelength array of eight aerials each fed with current in time-phase will be as shown in Fig. 175a, and it is observed that a perfectly symmetrical bi-directional polar diagram of radiation is produced, because the vectors wrap up either in a clockwise or an anti-clockwise direction from an initial straight-line vector both sides of the normal.

Let us now alter the time phases of the currents in the eight aerials of our typical array, so that the current in aerial 1 is leading by a small angle on that in aerial 2 and so on. This will result in the vectors having an initial bias because the time-phase reduces the field along the normal to the array line.

Considering that direction (to the right) where the space-phase between aerials is equal to the time-phase we have applied, the time-phase and space-phase will be in opposite sense because in this direction the space-phase of No. 1 vector is

lagging on No. 2 and so on, whereas the time-phase is leading. This will result in the vector sum being straightened again, thus giving maximum field for this direction. Contrariwise at an equal angle θ from the normal in the left-hand quadrants, the space-phase is additive to the time-phase and the field is reduced. The result of this small time-phase is therefore to give a bias to the maxima of the polar diagram to that side away from the leading time end and make it asymmetrical as shown in Fig. 175b.

If we now increase the time-phase to such an extent that it is exactly equal to the space-phase between elements (in this



a b c

FIG. 175. Polar Diagrams for Different Phases of Current.

case 90°) the vectors for directions normal are those of Fig. 173d (left) giving a zero resultant. In directions to the right the vectors unwrap themselves and just become unwrapped for a direction in line with the array as in this direction the space-phase is now equal and opposite to the time-phase. To the left the vectors become more wrapped up still and produce small tails. Thus the polar diagram is biassed to such an extent that the maxima now overlap and produce a uni-directional polar curve lying along the array line shown by Fig. 175c, the maximum of this polar curve being away from the end which is leading in time-phase, in this case No. 1 aerial.

If desired we can completely reverse this polar curve by changing the time-phase such that the opposite end (No. 8) is leading in time-phase, say, by feeding from the end No. 8 instead of No. 1. Alternatively, if we still desire to feed the system from the No. 1 end, we must misphase each succeeding aerial by 360 degrees minus the space-phase, as this is the same as giving the next aerial a lagging current.

We have considered the case of eight aerials each spaced $\frac{1}{4}\lambda$ apart. If we were to position our aerials more closely together we should obtain the same result if we altered the time-phase to equal the new spacing. Thus if our aerials are arranged to be one-sixth of a wavelength apart, i.e. 60°, we should need to produce a time-phase between radiating elements of 60° or $(360° + 60°)$ to produce a maximum away from the fed end ; or $(360° - 60°)$ to produce a diagram with maximum towards the fed end.

As explained previously the directive properties are not materially changed with spacing of elements, and in general a $\frac{\lambda}{4}$ spacing is customary.

Thus we have the general rule for a single line of radiators that with zero (and opposite time-phase) or multiples thereof, bi-directional symmetrical figures are produced, the former with maxima normal, and the latter with maxima in line with the array.

If the time-phase is made equal to the space-phase between radiators a uni-directional polar curve is obtained with a maximum away from the feed or leading time-phase end. With a time-phase equal to 360° minus the space-phase, the polar curve is reversed and the maximum now points towards the feed or leading time-phase end.

It is evident that there will be mutual impedances between each aerial and all the others and, if the aerials are to be in tune, it will therefore be necessary to adjust the electrical length of the aerials to suit the spacing. This adjustment will be done experimentally, with previous experience as a guide, since the problem would evidently be very cumbersome to solve theoretically.

The vector diagrams of Fig. 173 are based upon a uniform current distribution. Existing end-fire arrays are, however,

only fed from one end and hence the current tapers off (mainly due to radiation) as we proceed along the array, and this will influence the shape of the polar diagram.

Thus for the same array of eight aerials but assuming that the current in each aerial will be 0·875 of that in the preceding one, the equivalent vector diagrams will be as shown in Fig. 176. For the particular case of an array having a time-phase α equal

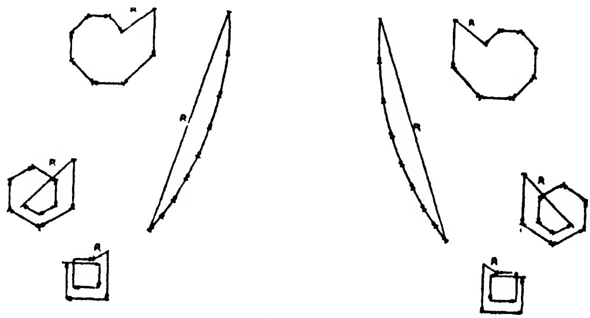


FIG. 176.

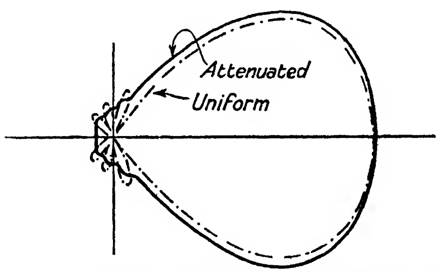


FIG. 177. Illustrating Effect of Attenuated Current.

to the space-phase (given previously in Fig. 175c), the polar diagram will now be as shown in Fig. 177 full line.

It is seen by comparison with Fig. 175c (shown again dotted in Fig. 177) that the shape of the main loop is but little changed. The radiation is not zero in any direction, but since the smaller tails are not so much in evidence the general efficiency is of the same order.

Broadside Arrays

It is clear that a single line of aerials, carrying currents in phase with each other, produce a bi-directional diagram. With

such broadside arrays it is therefore necessary to provide a second line of radiation in the rear, to reinforce the forward radiation and cancel the backward.

It is usual (because more convenient) to supply only the front

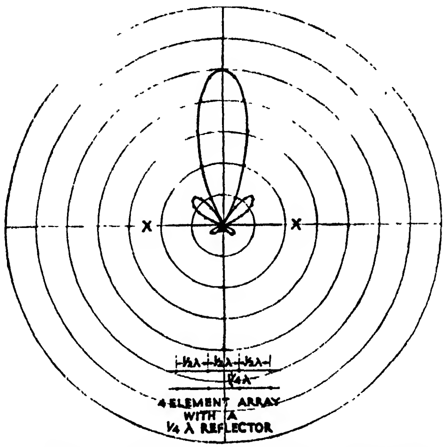


Fig. 178. Polar Diagram of Broadside Array with λ/4 Reflector Spacing.

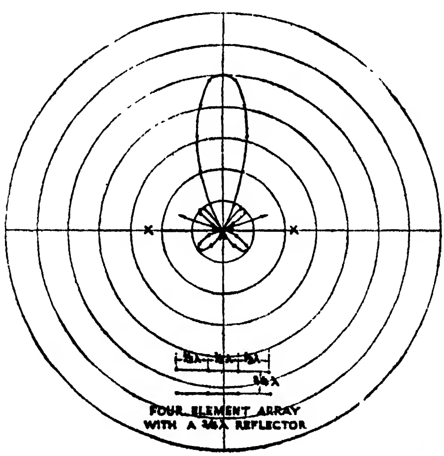


Fig. 179. Polar Diagram of Broadside Array with ⅜λ Reflector Spacing.

line and let the rear line act as "reflectors" or "parasites." Where we are concerned with lines of aerials it is found that a spacing of $\frac{\lambda}{4}$ is about the best to employ and the lengths of

the reflectors are adjusted to give the best ratio of forward to backward radiation.

The tuning and spacing adjustments which would cancel backward radiation are not the same as those which would increase forward radiation to the maximum extent and a compromise is therefore adopted.

In Figs. 178 and 179 polar curves for a 2λ array, having reflectors $\frac{\lambda}{4}$ and $\frac{3\lambda}{4}$ respectively behind the active aerials, are shown. They assume that the reflector currents are equal in magnitude to those in the front aerials but are in quadrature (leading).

From these figures it can be observed that the polar diagram is characterised by a main lobe and a number of side lobes or tails. It will be remarked that the wider the array, the more concentrated is the main lobe and the greater the number of tails. The width of an array, measured in wavelengths, is termed its aperture. An array built up vertically gives a concentration of energy in the zenithal plane in exactly the same manner.

In assessing the value of an array we must be careful to take into account its current distribution. For instance, an array of A square metres fed at one point (the centre, for example) would not be nearly so effective as the same array fed at, say, four points. With a centre-point feed only, there would be a tapering current to the edges and, in consequence, the effective area would be reduced.

If it is very important that the side lobes should be as small as possible (usually a requirement of arrays for radar or navigational systems), then it will be necessary to taper the currents because this reduces the magnitude of the side lobes, at the expense of widening the main lobe.

Array systems may be made to radiate either vertically or horizontally-polarised waves and there are many examples of both types of array in practice. Theoretically, as has been seen in a previous chapter, there should be little to choose between the two types of waves for long-distance work and both are about equally used.

We will now describe a few typical examples of broadside arrays.

Marconi Broadside Array

The desirable features of an array discussed in the general requirements paragraph can best be attained by means of an array of separate isolated unit radiators. As it is not economic to provide too many feed points, a system must be developed with some indirectly energised wires, which can be numerous without increasing complication.

Since the economical height of an array is limited and the zenithal polar diagram very important, it is a difficult but essential matter to provide as good a zenithal diagram as possible with a reasonable height. The Franklin uniform aerial

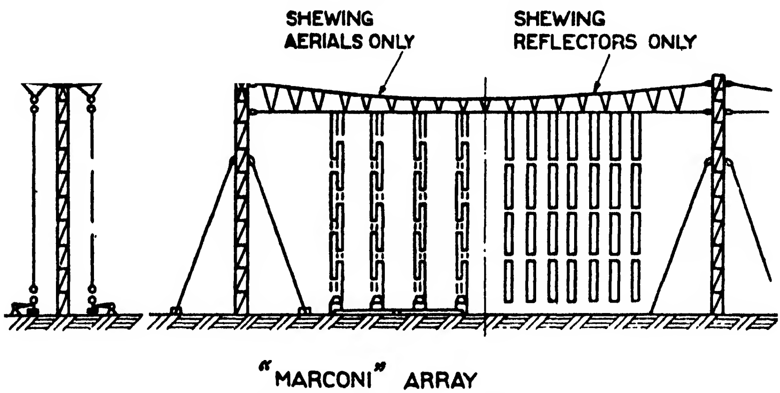


FIG. 180. Marconi Broadside Array.

(described on page 273) gives the best efficiency in this respect since it is the only aerial in which the phase-reversing device actually assists in forward radiation, and the aerial approximates very nearly to a uniform current sheet.

A sketch of a typical Marconi beam array is given in Fig. 180, aerials and reflectors being shown separated for clearness. The aerials are of the uniform type and the reflectors each comprise two, or sometimes three, parallel wires close spaced. The effect of increasing the number of reflector units is not to improve the diagram materially, but to flatten the tuning of the array. At the same time it is found this increase of reflectors does decrease backward radiation.

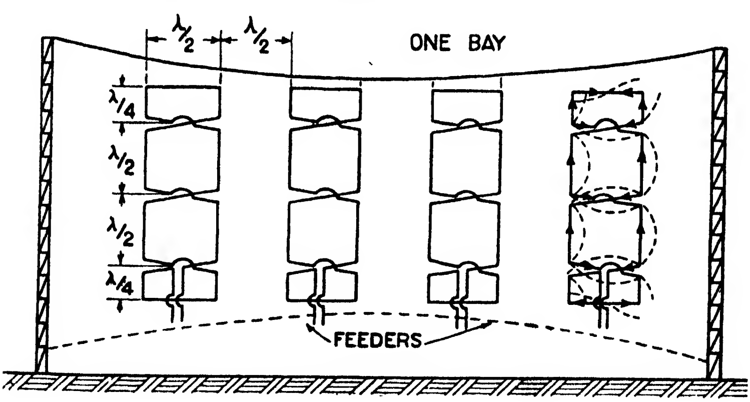
The reflector aerials are usually placed approximately a quarter of a wavelength behind the energised aerials in the

longer wavelength arrays, and three-quarters of a wavelength for the shorter waves, the lengths of individual aerials being adjusted to give the best diagram. There are twice as many reflector units as energised aerials, their spacing being approximately $\frac{\lambda}{4}$ whilst that of the energised aerials is $\frac{\lambda}{2}$.

The number of aerials and, therefore, the length of the array naturally depends upon the narrowness of the beam which it is economically or otherwise desirable to produce. With one or two exceptions eight wavelengths is the maximum aperture made, and six, four and two wavelength apertures are also used. The feeder system supplying the aerials in the same phase has already been discussed in Chapter VI.

The Sterba Array

The Sterba Array was developed in the U.S.A., and is used on a number of the telephone links constructed by the International Telephone and Telegraph Corporation. The array is



"STERBA" ARRAY.

FIG. 181.

built up of a number of units of the form shown in the Fig. 181. This unit consists of a wire which, when it is supplied with current at the correct frequency, will carry a system of stationary waves. The wire is bent up in such a way that half wavelengths, all of which are of the same phase, are vertical

($\theta^\circ = 10^\circ$) to the right. Then if A and B (or A' and B') be two points on a feeder system Y metres apart along a line normal to the wavefront, B must be given a lag by inserting

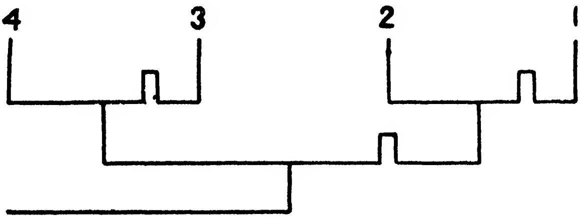


FIG. 183.

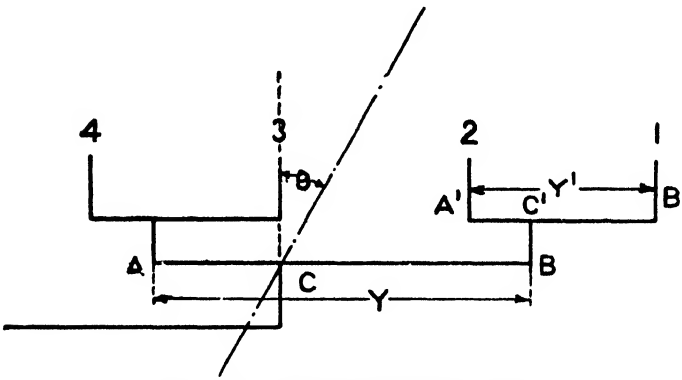


FIG. 184. Swinging a Broadside Array.

an additional length of feeder between C and B, the amount required being

$$X = Y \sin \theta$$

It is easy to understand how accurately phasing by such means can be accomplished, as even at the highest frequencies quite an appreciable length of feeder will represent only a degree or so of phase shift. For instance, if the wavelength is 20 metres, then 20 centimetres length will represent no more than 3·6°.

"End-Fire" Arrays

We have already shown theoretically that by making the time-phase of radiators in a single line equal to the space-phase, a maximum energy concentration will be produced along a

single array line without a reflector curtain, and in this case since the concentration in all planes is a function of length, the array height can be small, and the prime cost in consequence much less than the broadside array.

Although the " broadside " array gives a sharper, horizontal polar-diagram than the " end-fire " type for a given horizontal length, if the solid polar diagrams are compared, both may be equally effective in concentrating energy. In the same way that the maximum useful aperture of a broadside array is limited so the useful length of an " end-fire " array is restricted to an even greater extent and end-fire arrays are seldom made greater than four wavelengths.

Different forms of " end-fire " arrays are now in use, but in many ways the results obtained with them are disappointing compared with the broadside types for transmission purposes. It is possible this may be because of one or two reasons :

1. Most " end-fire " arrays are terminated, whereas broadside arrays are all of the stationary-wave type. Terminated arrays are generally inefficient for transmission purposes because a percentage of the total power into the arrays must of necessity be wasted in the terminating resistance.
2. All " end-fire " arrays are fed from a single feed-point at one end.

As regards the second feature, all " end-fire " arrays are fed from one end, as it is such an elegant solution to the problem of making the time-phase equal the space-phase. The authors are inclined to believe, however, that this solution, although elegant, is ineffective and that not until designers have the courage to adopt a multiple-feed system, which will level up the distribution of energy along the array line, will the end-fire array be able to compete with its broadside counterpart.

R.C.A. Arrays Using Long Radiating Wires[12]

We have seen that a long " harmonic " aerial directs its main radiation at a decreasing angle with its own axis as the length is increased.

For instance, the polar diagram of a wire eight wavelengths long is characterised by a main loop making an angle of $17\frac{1}{2}°$

with the axis and a number of small tails. The same diagram will be obtained for all angles perpendicular to the axis of the wire, and hence the solid polar diagram consists of two main cones of radiation concentric with the axis of the wire together with a number of subsidiary cones. This property of a long wire forms the basis of two arrays briefly discussed below.

Suppose we erect a wire A at an angle of 5° with the ground, then A, B, C, and D (Fig. 185) depict the zenithal sections of the main cones of radiation, the small tails being omitted. The maximum radiation from section A is then at an angle of $12\frac{1}{2}°$ with the earth's surface, this being considered the most

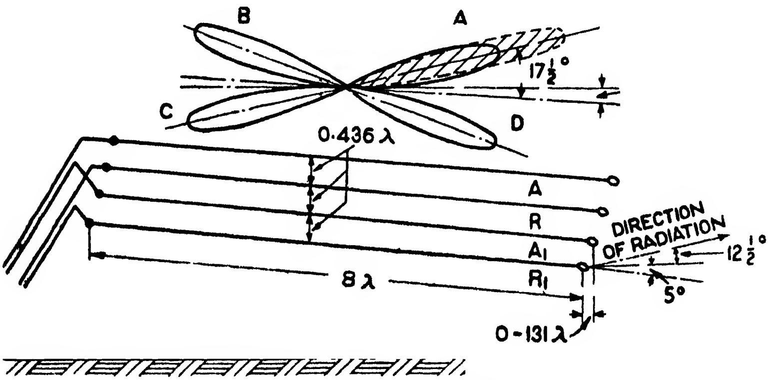


Fig. 185. R.C.A. Array.

useful zenithal angle to use for communication purposes on the particular system considered.

If a second wire A_1 is erected, parallel to the first, and fed in opposite phase, such as from a transmission line, the total field will fall away to zero in the plane normal to that containing both wires; because any distant point is equidistant from each, and the wires are fed in phase opposition. Thus the polar diagram shown is true only for the zenithal plane, and some measure of horizontal directivity has been obtained.

It is necessary still to get rid of three of the four main lobes from the diagram. It will be seen from Fig. 185 that the wires are staggered, the amount of displacement endwise being such that in a direction B, or D, any point of one wire is in phase opposition to the corresponding point on the other.

The effect of this is that in these directions a zero field is obtained, and these lobes are each split into two small tails. In order to get rid of the remaining lobe C, it is necessary to erect a second pair of wires R and R_1 supplied with current in phase opposition to each other, and in quadrature with the currents in A and A_1. When the current in R leads 90° on that in A, then lobe C breaks up into small tails and hence the required uni-directional diagram is obtained.

Another arrangement of the harmonic aerial is the " V " type shown in Fig. 186. The harmonic aerials are set at an angle and fed at the centre and thus the currents in A_1 and A_2 are in phase opposition with each other, and it follows that the cones of radiation produced by each wire at an angle of $17\frac{1}{2}°$ with itself will unite if the wires are folded at an angle of 35°. The maximum radiation will, therefore, be along the line bisecting the " V," but the arrangement will be bi-directional. A similar " V " is therefore erected, an odd quarter wavelength behind the first (usually $2\frac{1}{4}\lambda$), and supplied with current in quadrature, either leading or lagging, according to whether transmission in directions RA or AR is required.

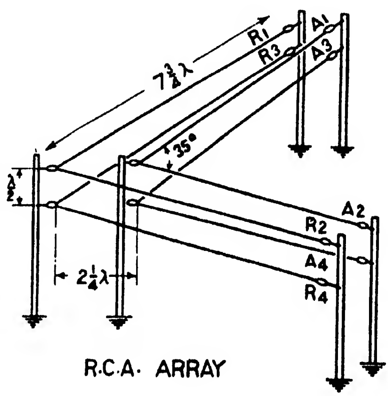


FIG. 186. R.C.A. " V " Array.

By the use of a second " V " a half wavelength below the first, the high angle radiation is reduced. It will be seen that vertical radiation would be entirely cancelled.

Both the above arrays produce, of course, horizontally polarised waves.

Terminated " End-Fire " Arrays

Consider a vertically-polarised, horizontally-propagated wave arriving at a vertical wire connected at its base to a receiving system whose input impedance is equal to the equivalent

impedance R_0 of the wire, as shown dotted in Fig. 187. It will be assumed that the far end of this wire will also be terminated in a resistance equal to R_0 and therefore in our discussion we shall not consider reflection from the top end of the wire as affecting the problem. The arriving wave will induce in each elementary length of wire an E.M.F., and a current wave will therefore be initiated from each element, which will travel down the wire to the receiver input. The problem can virtually be considered as a feeder collecting energy along its length and the resultant E.M.F. at the receiver input will be dependent upon the space-phase of induced E.M.Fs. plus the time-phase due

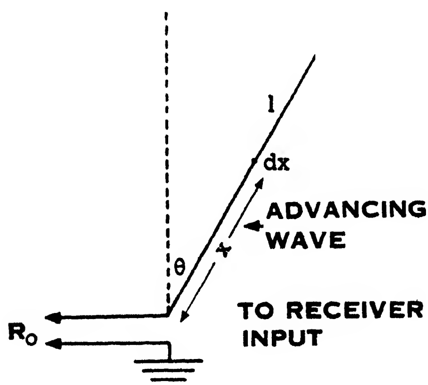


FIG. 187. Terminated, Tilted Wire.

to the feeder length from each element. In the case cited the space-phase is zero and the time-phase dependent upon the length of wire. For instance, if the wire is $\frac{\lambda}{2}$ long the space-phase is 0°, but the time-phase is such that the wave from the topmost element will lag 180° on that from the bottom element and hence the vector diagram will form a semicircle. Whereas if the wire is λ long the vector diagram will be a circle.

If now we advance the top end of such a wire into the wave by tilting the wire (as shown full line), we shall introduce a space-phase which will advance the induced E.M.F. in the topmost end of the wire to the greatest intent and the resultant vector diagram will unwrap itself. Assuming for the moment

that tilting the wire does not reduce the E.M.F. induced in each element, the more we tilt the wire the more influence the space-phase has and tilting to a horizontal position will give the maximum resultant E.M.F. to the receiver, because in this position the space-phase exactly equals the time-phase and the resultant vector diagram has thereby been unwrapped from whatever shape it was with the wire vertical to a straight line ; this is true no matter what the length of wire.

Conversely, tilting the wire away from the direction of the advancing wave will wrap the vector diagram up more and more and in consequence the system will be seen to have directional properties, that is assuming there is no reflection from the far end. From previous sections it will be clear that the polar diagram will be dependent upon wire length, the longer the wire the better the diagram.

The tilting of a straight wire will not, however, be very efficient, because as it becomes more and more horizontal each element will have reduced radiation efficiency for vertically polarised waves (reception efficiency in the case of a receiver wire), the zenithal polar curve of each element having a cosine law.

R.C.A. Long Wave Array

Beverage used such a horizontal terminated wire (Fig. 188) as a receiving aerial for long wavelengths at Riverhead, U.S.A. This aerial, which was 10 miles in length, had good directional properties but poor radiation efficiency for reasons explained above ; in fact, if the received waves were exactly vertically polarised no E.M.F. would be received by such a wire, but at Riverhead the soil is very dry and sandy and this had the effect of giving the received wave a considerable forward tilt so that a horizontal component was in evidence which induced E.M.Fs. in each element of the wire. The terminating resistance R_0 shown in Fig. 188 prevents reflection from the far end and gives the array unidirectional properties, because although waves arriving from opposite directions will build up a large E.M.F. at R_0, the energy there is completely absorbed and not reflected back to the receiver end.

R.C.A. Fishbone Array

Such an array would appear to be as suitable for short waves as for long, but since short waves are variously polarised, a

modified form of array has been built consisting of a two-wire transmission line of low surge impedance (about 350 ohms) to which horizontal pick-up wires are coupled as shown in Fig. 188,

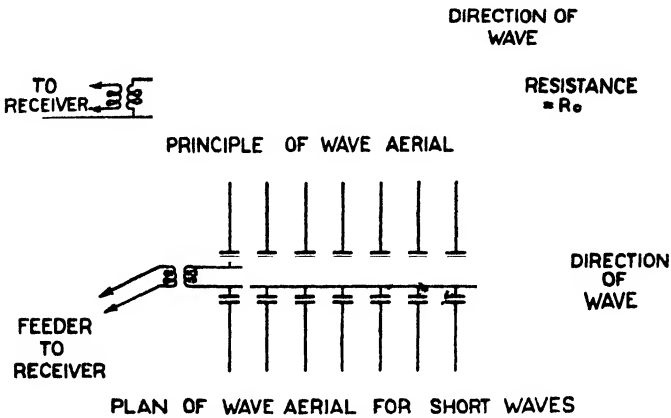


Fig. 188. R.C.A. "Fishbone" Array.

and as with the original Beverage aerial, the far end is terminated by a resistance R_0 to produce a unidirectional diagram.

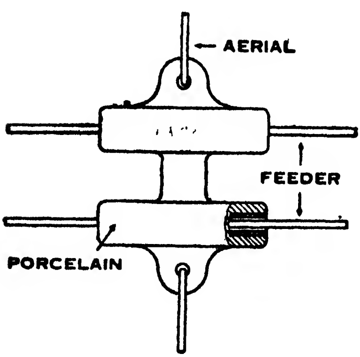


Fig. 189. "Fishbone" Array—Coupling an Aerial to Feeder.

As produced by the R.C.A. this array was designed to have flat tuning, covering a wave-range of about 4/1 efficiently, this being accomplished by using "pick-up" wires. If λ is the shortest wavelength to be received, then the spacing is $\frac{\lambda}{6}$ and the length of each wire rather more than $\frac{\lambda}{4}$. The wires are coupled to the transmission-line through capacitors of special design (Fig. 189), and because of this small capacity in series with the wires they are electrically less than $\frac{\lambda}{4}$ and present a capacity reactance to the transmission-line at all wavelengths being received. Because the velocity of the wave along the

line is lower than that of the arriving wave in space, it is not useful to make the array more than 8λ long. For a greater length, the voltages contributed to the receiver by the furthest wires would lag so far behind those from the nearest ones as to add but little to the vector resultant.

Messrs. Cable and Wireless, Ltd., have used a modified form of this array for spot-frequency working by eliminating the coupling condensers, lengthening the pick-up wire to $\frac{\lambda}{2}$ (less 25%) and increasing the spacing between wires to $\frac{\lambda}{4}$. Since they are end-fed to the feeder line this is spaced more widely, so as to bring its characteristic impedance to 600 ohms. The array length is also reduced to about 4λ and when increased gain is desired, arrays will be paralleled.

Rhombic Arrays

Returning to a consideration of the tilted wire coupled to a receiver of input impedance R_0 as shown in Fig. 187, it was mentioned that as we tilt the wire more and more into the wave, although the space-phase is tending to counteract the misphase due to the wire length, the pick-up in each elemental length is getting smaller. From our previous discussion on array systems it is easy to derive an expression for the relative E.M.F. produced at the receiver for different angles of tilt, treating the wire as a feeder line to which is coupled an infinite number of elemental pick-up aerials.

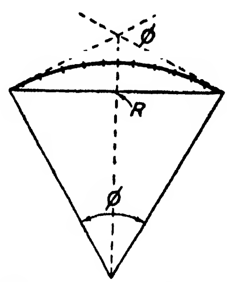


FIG. 190. Vector Diagram for Tilted Wire.

Let the field strength be E volts per metre. Dividing the aerial up into a series of elemental lengths dx gives us a series of elemental generators of voltage $E \cos \theta \, dx$. These generators are not in phase with respect to the current in the base impedance, but the phase between them is determined by summing up the phase lag given by the time taken to travel from the given element position to the base and the lead (or lag) imparted by tilting the aerial into (or away from) the source of transmission.

This condition is represented by a series of vectors of length $E \cos \theta \, dx$ distributed round the circumference of a circle, as shown in Fig. 190, page 323. Now the total phase angle ϕ between the first and last elemental vector is equal to the angle due to the time of travel along the wire from B to A minus the angle of lead due to the tilting of the wire into the wave. That is

$$\phi = \frac{2\pi l}{\lambda} - \frac{2\pi l}{\lambda} \sin \theta \quad . \quad . \quad . \quad . \quad (9)$$

The resultant voltage at A is represented by the vector R, and as has been shown previously this subtends an angle ϕ at the centre of the circle of vectors. Thus :

$$R = 2r \sin \frac{\phi}{2} \quad . \quad . \quad . \quad . \quad . \quad (10)$$

where r = radius of circle.

The arc $= r\phi = l\,E \cos \theta$

or $$r = \frac{l\,E \cos \theta}{\phi}$$

$$\therefore R = \frac{2l\,E \cos \theta}{\phi} \sin \frac{\phi}{2}$$

$$= \frac{l\,E \cos \theta \sin \frac{\phi}{2}}{\frac{\phi}{2}}$$

$$= \frac{l\,E \cos \theta \sin\left\{\frac{\pi l}{\lambda}(1 - \sin \theta)\right\}}{\frac{\pi l}{\lambda}(1 - \sin \theta)} \quad . \quad (11)$$

If this expression is differentiated to find the angle at which the resultant R is a maximum, it will be found that for each length of wire there is an angle of tilt which gives a maximum resultant from a vertically-polarised wave and it is clear from the previous discussion that the longer the wire the greater the tilt angle necessary to obtain a maximum pick-up,

the curve connecting tilt angle and pick-up being shown in Fig. 191. Observe that for short wires the tilt angle is critical,

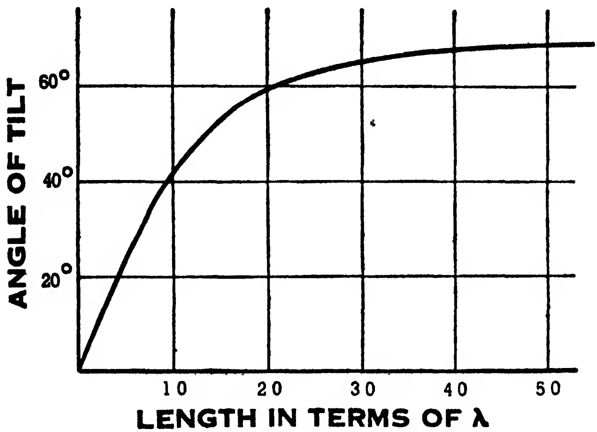


FIG. 191. Optimum Tilt Angle.

but less so for long wires. Since the foregoing only presupposes travelling waves we must terminate the far end by a system

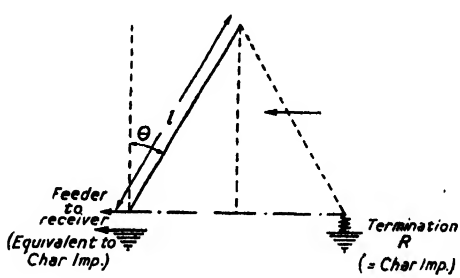


FIG. 192. Tilted-wire Aerial.

which will not reflect, say, by adding a second wire (shown dotted in Fig. 192), terminated in a resistance $R = R_0$.

Rhombic Array. International Telegraph and Telephone Company

This principle has been applied in the well-known Rhombic (or Diamond) aerials first produced by the I.T. and T. Company, but they are usually constructed for the reception of horizontally-polarised waves by turning the wire system to a

horizontal position and adding a second pair, as shown in Fig. 193. This setting up of a horizontal rhombus at a height h above earth will, of course, modify the polar diagram and the value of received voltage from a given field, and by small alteration of the wire lengths and angle it is possible to arrange for maximum directivity to be at any given angle β from the horizontal.

An interesting and useful feature of the Rhombic aerial,

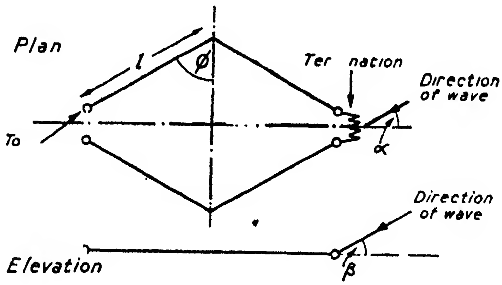


FIG. 193. Rhombic Array.

which is of particular value in reception work, is that the aerial has a broad tuning characteristic and is therefore suitable for operating over a wide wave-band. This will be evident from the consideration of the diagram showing the angle of tilt, which changes but slowly for long wires such as would be used. For a normal array, a 30% change of incoming frequency would only produce a drop of some 2 db in the input to the receiver.

Rhombic Transmitting Aerials

The single rhombus aerial just discussed is really only suitable for reception purposes because of its low radiation efficiency. If used for transmission purposes, some 40% to 50% of the power input would be wasted in the terminating resistance. It is found, however, that rhombics may be grouped in series or parallel in such a way that their combined directivity is maintained (or even improved) for a given area occupied by the complete array and the radiation efficiency can be brought up to more than 90%. Various forms of such grouped rhombics have been evolved by the Marconi Company, three of which

are shown in Figs. 194 and 195, the arrangement of Fig. 194 being the form most used. As with a single rhombus, the

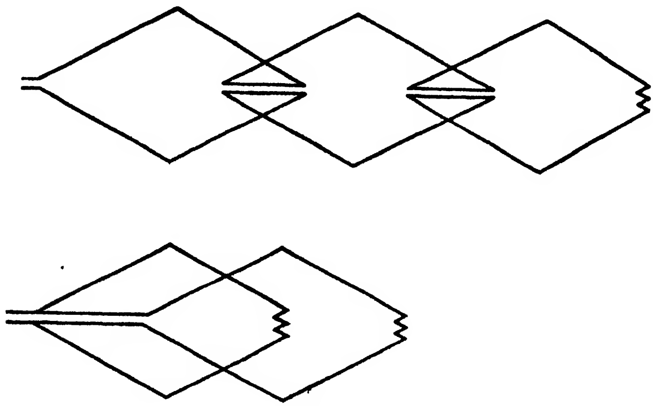


FIG 194.

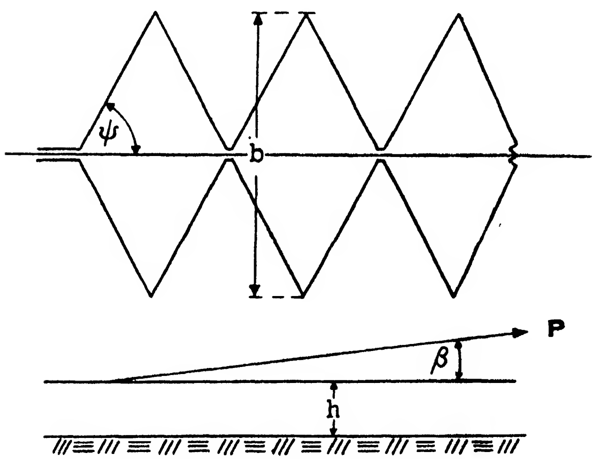


FIG. 195. Modified Rhombic Arrays.

dimensions can be arranged so as to obtain a maximum directivity at any angle to the horizontal. It should be remarked, however, that such a system is now critical as to wavelength.

Marconi Series-phase Array

A form of "end-fire" terminated array, devised by C. S. Franklin and known as the Series-phase Array, contains some

interesting features, and its simplest form is shown schematically in Fig. 196. It consists of a wire folded into a number of loops which act as radiators, connected by horizontal wires.

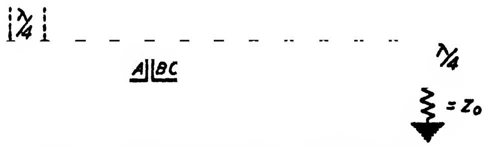


FIG. 196. Marconi Series—Phase Array.

The dimensions of the loops and spacing pieces are determined by the diagram required, 12 to 16, $\frac{\lambda}{4}$ loops, spaced $\frac{\lambda}{4}$ apart, being the most common arrangement, thus making an array three or four wavelengths long.

Each loop acts as a $\frac{\lambda}{4}$ aerial, as can be seen from the following considerations. If we have two equal current waves travelling in opposite directions on the same wire, then we know that the result is a stationary wave. In the series-phase array the currents are in two separate wires but these are so close together that the radiation reaching a distant point will be the same as that from a single wire carrying a stationary wave.

If the currents at adjacent points on the two wires are added (with due regard to their directions) it will be seen that the resultant is always zero at the top of the loop but is a maximum at the bottom.

The radiation resistance, referred to the current I flowing in the wire, will be four times that of a $\frac{\lambda}{4}$ aerial, because the equivalent $\frac{\lambda}{4}$ aerial is carrying a stationary wave having a value $2I$ at the base.

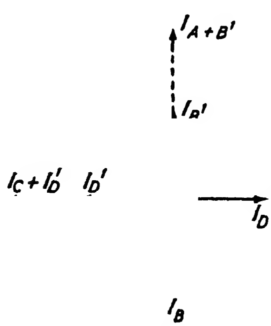


FIG. 197. Vectors for Series-Phase Array.

In order to determine the polar diagram, we need to know the relative phases of the radiation produced by the equivalent

$\frac{\lambda}{4}$ aerials. If we take the vector for the current I_A at A as the reference, then I_B will lag 180°, I_C 270° and I_D 450° as shown in Fig. 197.

To find the phase of the equivalent stationary wave in AB, we reverse I_B, since it is flowing in the opposite direction, and add to I_A. Similarly, we add the reversed I_D to I_C to obtain the stationary wave in CD. It will be seen that this leads 90° on that in AB. Hence along the line of the aerials, *looking towards the fed end*, the time-phase and the space-phase cancel out and the maximum of the polar diagram will be in this direction.

If similar vector diagrams are drawn for the case of $\frac{\lambda}{2}$ loops, spaced $\frac{\lambda}{4}$ apart, it will be seen that maximum radiation is away from the fed end, but the more usual arrangement is that shown in Fig. 196.

The vector diagrams assume that the velocity along the wire is the same as in space. As this is not quite true in the actual array, the spacing and length of the loops needs to be adjusted slightly, to allow for this effect.

Since the array gives a good polar diagram and has a high radiation resistance, it is suitable for both transmission and reception but only over a limited frequency band. The currents naturally decrease as we proceed along the array but by varying the dimensions the current can be made more uniform and the frequency band increased, at the expense of efficiency.

The terminating resistance will be made equal to the characteristic impedance of the wire, which, for the $\frac{\lambda}{4}$ loops, is about 300Ω.

Energy Gain of Arrays

If the solid polar diagram of a directional array is known and that of a single aerial is also known, then by integrating these polar diagrams we can find the power which each would use in order to produce the same field strength in the required

direction, and this power ratio (expressed usually in decibels) will be the gain of the array over the single aerial.

Franklin estimated originally that the energy gain of a broadside array system would be 9·6 per square wavelength of aperture surface compared with a half-wave aerial. The easiest comparison to make is to take one plane at a time and compare the gain of the array with one of the aerials which go to make it up, as we may suppose in this case that the vertical polar diagram is the same for both, since the same aerial height and quality of earth is involved.

T. L. Eckersley, Green, and Southworth, treating specific

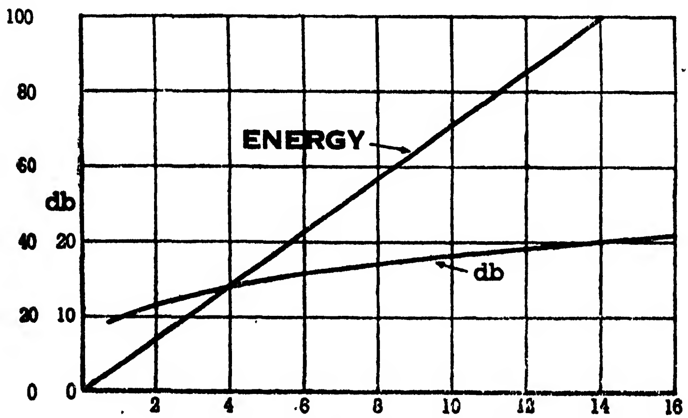


FIG. 198. Energy Gain of Arrays.

cases, have produced results which are in fair agreement with Franklin's, and the curve of Fig. 198 gives average values for the gain of an array of different apertures.

It will be seen from this diagram that the energy gain is directly proportional to the array aperture. Thus a 6λ array has an energy gain of 43 (16 db) and a 12λ array a gain of 86 (19 db). Accordingly, if we add a 6λ array to an existing array of 6λ, we shall double the energy gain. If, however, we have an array at both transmitting and receiving ends, the total gain of the system is now the product of the array gains, not the sum. This can best be seen by referring to the level diagram in Fig. 199. Consider an output power of 1 kW, i.e. 60 db above a datum level of 1mW, and an ionosphere attenua-

tion of 150 db. With no arrays, the level of received signal will be − 90 db, shown by curve 1. If we use at the transmitter an array having a 16 db gain, we raise the level at the transmitting end, and therefore at the receiving end, by 16 db,

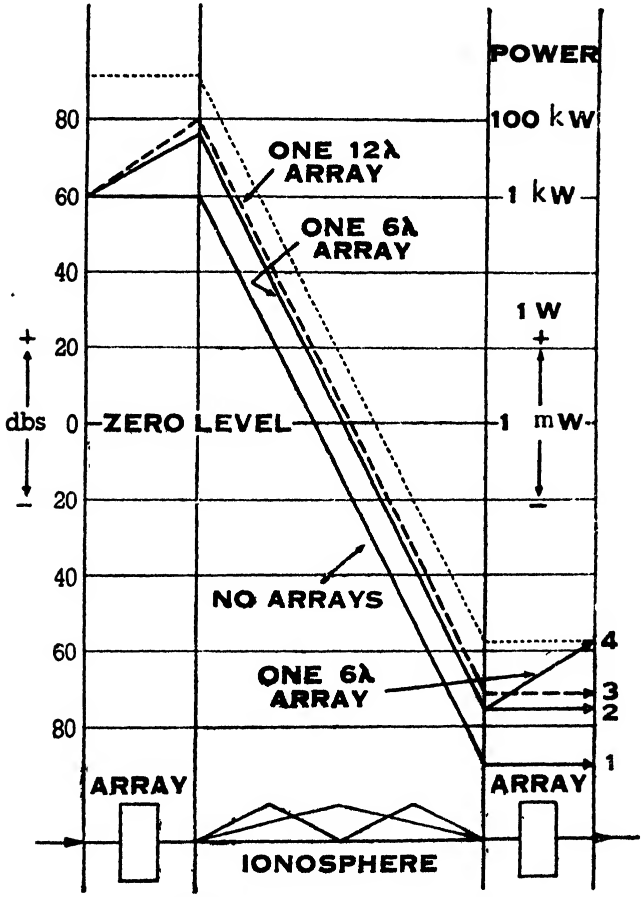


FIG. 199. Illustrating Overall Gain Due to Aerial Arrays

as shown by curve 2. Had we increased the dimensions of the transmitting array to 12λ, this would have increased the level at the transmitting and receiving end only by another 3 db, as shown by curve 3. On the other hand, had we left the dimensions of the transmitting array at 6λ, raising the level by 16 db,

and used a 6λ array at the receiving end, the output level to the receiver would be raised by another 16 db, i.e. the total gain would have been 32 db. If omni-directional aerials had been used at both ends, it would have been necessary to raise the transmitter power to 1,200 kW, to obtain the same level of

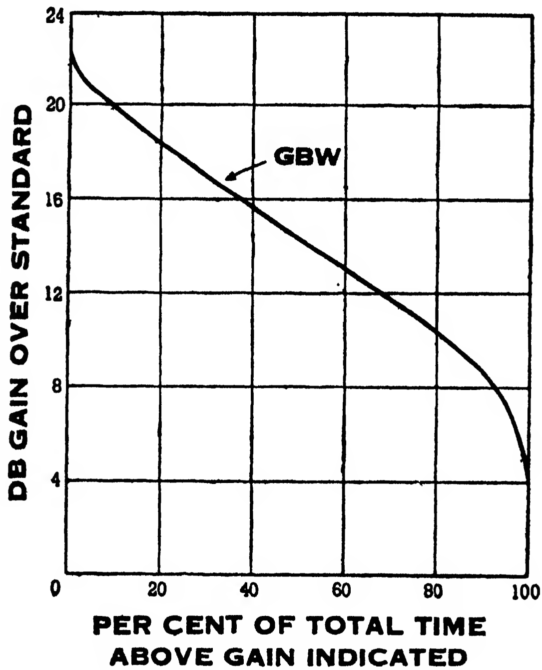


FIG. 200. Effect of Propagation Conditions on Gain of Array.

received signal, as is shown by tracing back the dotted line on Fig. 199.

It might be expected that calculated values for the gain of an array system would differ considerably from the experimental figures since a number of uncertain factors are involved, but surprisingly enough calculated and measured gains are in very fair agreement.

It is found, however, that the effective gain of an array does not remain constant but depends upon ionosphere conditions. This is because the full gain can only be obtained when we deal

with a perfectly-uniform wave front, as it is only then that the phase relationships in the various aerials are correct. Measurements taken over long periods indicate that only for a small percentage of the time is the full gain of the array obtained, Fig. 200 showing the average of some results made by Bruce on short-wave transmitter in Great Britain, as received in America.

The Steerable Antenna (M.U.S.A.)

It has been seen that most arrays are designed to provide a fairly-sharp zenithal polar-diagram having a maximum at a small angle from the horizontal in order to receive mainly the

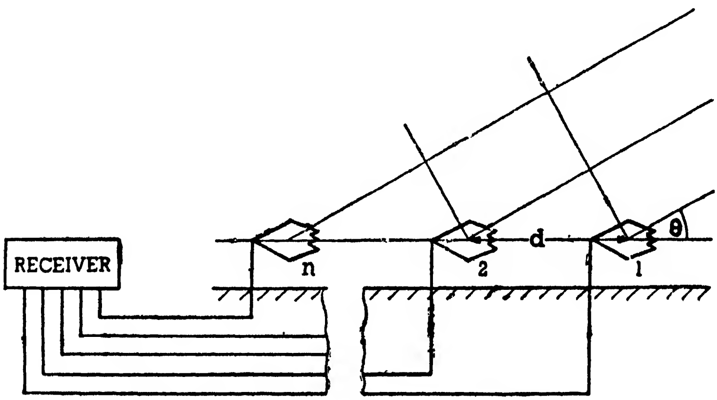


FIG. 201. Schematic Diagram of M.U.S.A.

low-angle rays which have made the fewest " hops " and which are usually the strongest. The disadvantages attendant upon reception of too many of the rays have already been discussed (see page 152). It is evident that if the zenithal polar-diagram is to be fixed it cannot be too sharp, since the rays which are most prominent and useful vary in their angle of arrival. The M.U.S.A. system, which has been developed by the Bell System Laboratories and is in use at both ends of the transatlantic telephone service, employs a long line of horizontal, rhombic arrays and is capable of very great directivity in both planes. The vertical polar diagram is, however, instantly varied by electrical adjustments at the receiver and may therefore be adjusted to suit the varying conditions.

An outline diagram of the arrangement is shown in Fig. 201, it being noted that the receiver is located in a line away from

the incoming signals. Let us concentrate our attention upon a ray arriving at some angle θ. This will arrive at (2) at a time $\frac{d \cos \theta}{c}$ secs. later than at (1) (where c is the velocity of the wave in space). The distance along the feeder from (2) is d metres less than from (1), however, and hence, if the velocity of the wave along the feeder is v, the voltage at the receiver provided by (2) will lead on that from (1) by $\frac{d}{v} - \frac{d \cos \theta}{c}$ secs. and the phase angle between the voltages is therefore

$$2\pi f \left(\frac{d}{v} - \frac{d \cos \theta}{c}\right) \text{ radians.}$$

Similarly, the voltage from the nth aerial will lead on that from (1) by

$$2\pi f \left\{\frac{d}{v} - \frac{d \cos \theta}{c}\right\} (n - 1) \text{ radians.}$$

If we now provide phase-shifting arrangements in each feeder at the receiver end, we can compensate for the phase differences so that all the voltages add up in phase and produce the maximum possible resultant. This compensation will, however, only apply to the ray at the angle θ and hence this ray has been selected from others which may be present, and evidently, by altering the phase-shifts, any ray can be selected. Two or more separate phase-shifting arrangements can be provided so that several rays can be selected and later combined, though further phase adjustment will be necessary before combination because the different rays will have travelled different distances.

It will be seen that the amount of phase-shift required depends upon the frequency of the signal, in the simple system we have described. For this reason and also because it would be very difficult to control the phase-shift at short-wave frequencies, in the actual equipment used the phase-shifting is carried out in a later stage of the receiver where the frequency is fixed. The method used is described in outline on page 631.

Arrays for Ultra-short Waves

The types of array which have been discussed can also be used for waves from, say, 1·5 to 10 metres, as well as for short

waves, and it naturally becomes easier and cheaper to produce a very sharp polar diagram as the wavelength is reduced. Some additional types are also considerably used on wavelengths of 1 and 2 metres.

Below 1 metre there would be great difficulty in carrying out the adjustments for correct phasing and termination required with some of the short wave arrays. Losses in, and radiation from, any complicated feeder system are also likely to be serious. Owing to the short wavelength, designs of arrays which are more on optical principles become practicable and are much used.

Yagi Array

This type of end-fire array, developed in 1928 by Yagi, has been much used on ultra-short wavelengths. The Yagi array has only one fed aerial, the others being "parasites."

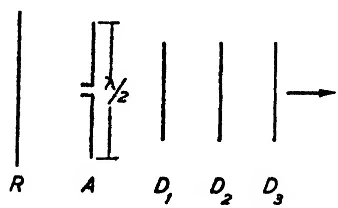


FIG. 202. A Yagi Array.

A typical array is shown in Fig. 202. A $\frac{\lambda}{2}$ aerial, A, is connected to transmitter or receiver. D_1, D_2 and D_3 are made somewhat shorter than $\frac{\lambda}{2}$ and therefore become "directors" (see page 300). R is made longer than $\frac{\lambda}{2}$ and is therefore a "reflector."

The calculation of the polar diagram is, in principle, the same as for the two aerials discussed on page 300. Each aerial has a coupling with each of the others, however, and a whole set of equations therefore results. The Mallock calculating machine has been used to obtain solutions for a number of arrangements.

Figs. 203, 204 show the horizontal and vertical polar diagrams (derived from curves given in Walkinshaw's paper) for an arrangement of one driven aerial, one director and one reflector. Two different values of reactance are shown and it will be seen

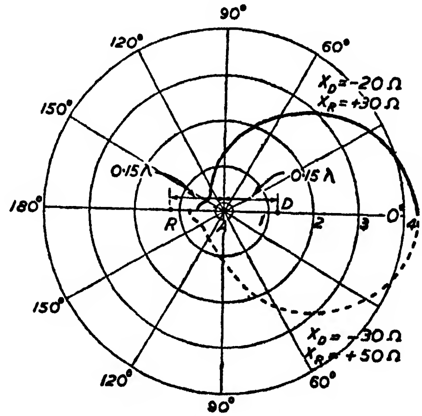


Fig. 203. Horizontal Polar Diagram of Yagi Array.

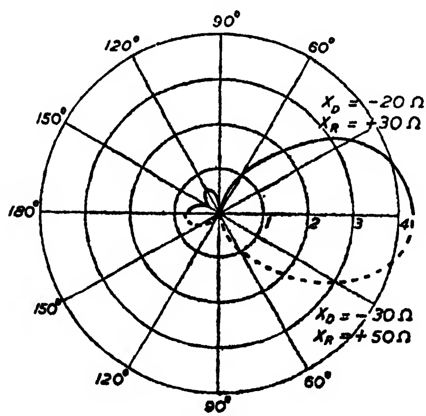


Fig. 204. Vertical Polar Diagram of Yagi Array.

that the polar diagram does not vary rapidly with change of reactance. The aerials are supposed to be in "free space," that is, the effect of the earth is not taken into account, but on the ultra-short waves, for which this aerial is likely to be used, the aerial will usually be a number of wavelengths above earth.

It is not found that the use of more than one reflector is effective, as the second reflector carries only a small current and therefore has a very slight effect upon the polar diagram. A number of directors can, however, be usefully employed with a sharpening up of horizontal and vertical polar diagrams.

For some arrangements the input resistance of the driven aerial tends to get inconveniently small. (The example on page 301 suggests this.) A folded $\frac{\lambda}{2}$ aerial is sometimes used, for which the resistance will be four times that of the ordinary type.

In place of a reflector, a wire screen is sometimes employed, placed about $0{\cdot}1\lambda$ behind the driven aerial.

Slot Arrays

The behaviour of slot aerials has already been briefly discussed and it is evident that an array can be formed from correctly-spaced slots, suitably fed, the most practical arrangement being formed by cutting $\frac{\lambda}{2}$ slots in the walls of a waveguide.

Such arrays fall into two main classes, depending upon whether the guide is closed by an adjustable piston (so that it is carrying a large stationary wave) or whether it is correctly terminated (and is therefore carrying a travelling wave). In the former case the phase of the wave will be the same all along the guide, except that it will be reversed in sign at each half-wavelength. In the latter, on the other hand, there will be a progressive change of phase depending upon the phase velocity in the guide.

For example, we can cut slots on the broad face of a guide, one-half of the guide wavelength apart, and displace each alternate slot on opposite sides of the centre-line. If the guide is carrying a stationary wave, then the amplitude of the guide wave at each slot will be the same but the phases at adjacent slots will be opposite. The spacing on alternate sides of the centre-line will, however, correct for this and we have, therefore, a line of radiators all of the same phase. The arrangement is, therefore, a broadside array. Like other arrangements

using stationary waves, it will only work satisfactorily over a small band of frequencies.

If a travelling wave in the guide is used, then the spacing of slots is usually about 200 electrical degrees (instead of 180°) at the mean frequency to be used. Because the phase velocity within the guide is different from that in space, the beam formed makes an angle with the axis of the guide. This arrangement can be used satisfactorily over a wider frequency-band. Since the wave in the guide is being attenuated, it will be necessary to cut the slots so that the coupling of the slots further from the feed point of the guide is greater.

Any of the slots mentioned on page 291 can be used to form an array. In all cases there will be coupling between the slots, just as between the aerials of other arrays, and this complicates the design.

Electromagnetic Radiation from Horns

If a waveguide is carrying an $H_{1,0}$ wave, for example, then at a distance from the open end large compared with λ there will be a radiation field of the same character as that produced by a dipole having its axis in the b direction. The radiation will be directed into a fairly sharp beam, the larger the dimensions the sharper being the beam, and vertical and horizontal polar diagrams are adjustable independently.

The polar diagrams depend upon the type of wave carried by the guide, which can be understood if we consider the analogy with a set of dipoles at the aperture of the guide. In the $H_{1,0}$ case these are all in phase, though having different amplitudes, but the higher-order waves would correspond to dipoles having different phases. Thus the higher-order waves may produce zero radiation along the axis and two or more principal beams.

If the end of a guide is used in this way, then a large cross-section is necessary in order to produce a sharp beam and there is considerable reflection back into the guide from the open end, because the effective impedance of free space is quite different from that of the guide. To radiate a greater proportion of the available energy and to direct it into a sharper beam, a horn may be used. This performs the same

directing and transforming action as an acoustical horn and will be the same in construction.

To obtain the sharpest possible beam for a given aperture of horn, the field at the aperture should all be of the same phase. but this will not be the case in a simple horn, unless it is many wavelengths long. The phase-shift may be corrected by placing

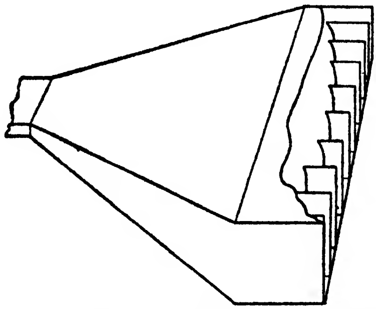


FIG. 205. Phase Correction in a Horn Radiator.

metal plates in the mouth of the horn (Fig. 205) so that the phase velocities (as in a waveguide) are increased over varying lengths and the phase at the aperture thereby adjusted.

Conical horns are an obvious arrangement for radiating from

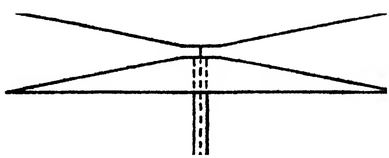


FIG. 206. Bi-conical Horn.

circular waveguides. When " all-round," horizontal radiation is desired, the bi-conical horn is a good solution (Fig. 206).

Parabolic Reflectors

The first short wave " beam " was produced by placing an aerial at the focus of a parabolic cylinder formed of vertical wires, but the arrangement was soon abandoned in favour of more efficient arrays, such as have been discussed. For waves of centimetre order, however, the parabolic reflector becomes a very suitable means for producing a sharp beam. In discussing

the behaviour of the parabolic mirror, we naturally think in terms of optics, though there are some modifications when dealing with radio waves.

It is well known that if a source of radiation is placed at the focus of a parabolic mirror, a parallel beam of radiation emerges from the aperture. This is only strictly true of a point source, in which all the radiation comes exactly from the focus of the parabola.

When using light waves of wavelengths such as $0{\cdot}5 \times 10^{-4}$ cm. it is necessary to accurately shape the mirror and to give the surface a high polish. When using centimetre waves, however, it is quite unnecessary to polish the surface and the mirror can, in fact, be built up of wire mesh or perforated material. This is useful when it is desired to reduce the windage and weight of the mirror.

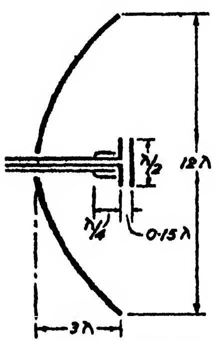


FIG. 207. Parabolic Reflector (not to scale).

The ordinary light source usually illuminates the whole mirror fairly uniformly, with unpolarised light of many different wavelengths. The aperture of the mirror is so many wave-lengths wide that direct radiation from the source has little effect upon the polar diagram.

In the radio case, however, the source may have quite a sharp polar diagram of its own and is producing polarised radiation, usually of one frequency or a narrow band of frequencies. Also, since the mirror will have a much smaller aperture (measured in wavelengths) the direct radiation from the source may be by no means negligible and will tend to spoil the polar diagram. Due to the polar diagram of the source, it will usually be necessary to use only a small portion of a parabola, so that the source is outside the aperture.

A typical arrangement, suitable for 10 cm, is shown in Fig. 207. In this case the focus is on the aperture plane and a parasitic aerial is used as a reflector, to direct as much as possible of the radiation from the aerial on to the mirror. The aerial is fed from a concentric line with a balancing arrangement such as that discussed on page 288.

The polar diagram in the plane shown in the sketch has a

width of about 8° (to half-power radii), whilst in the plane at right angles the width is about 6°. This difference is due to the cosine polar diagram of the aerial in its plane, which results in the mirror being less uniformly illuminated (and therefore less effective) in this plane.

When using centimetric waves the feed from the transmitter is often through a waveguide, rather than a concentric line. In such a case the aerial may be replaced by the open end of a waveguide, usually flared out to illuminate the mirror most effectively.

CHAPTER X

POWER AMPLIFIERS

SHORT wave valve transmitters may be divided into two types, self-oscillators, and driven circuits, or power-amplifiers as they are often called. The self-oscillator is but seldom used on short wave communication circuits except for small transmitters, and the power amplifier is now almost universal. The general principles on which power amplifiers are built are much the same whatever the frequency, but there are, in short wave working, several features that call for special attention.

The object of a transmitter is, of course, to produce high frequency power, modulated in accordance with the signal to be transmitted. The power radiated should be of constant carrier frequency and all the modulation frequencies should be reproduced in their correct relationships. Since the power involved may be considerable, we are also interested in the power efficiency of the transmitter.

A transmitter will generally include the following features:

(1) A driving source of constant frequency. This may be of the same frequency as finally radiated or an exact fraction. In the latter case, a series of frequency-multiplying stages will be necessary and will usually form part of the master-oscillator unit proper. These multiplying stages may or may not amplify as well.

(2) A chain of amplifier stages employing triodes, tetrodes or pentodes, working at the frequency to be radiated, each succeeding stage being of increasing power.

(3) Methods for stabilising the various stages. With triodes this will take the form of some balancing system to eliminate the feedback of energy through the inter-electrode capacities of the valve. With tetrode and pentode valves such a precaution generally becomes unnecessary, although circuit design requires care.

(4) The radio-frequency line, coupling the final amplifier to the aerial load.

(5) The keying or modulation system with its attendant " buffer " or isolator stage, to prevent carrier scintillation.

Leaving for the moment the driving source, the frequency-multiplying stages, circuit neutralisation and modulation, all of which are dealt with on other pages, one stage of amplifier proper will be considered.

Such an amplifier may employ a single valve, or a pair in push-pull (see page 375). The load on all stages except the last will be the grid input losses of the succeeding stage, whilst the final stage will deliver power into the aerial, either directly or

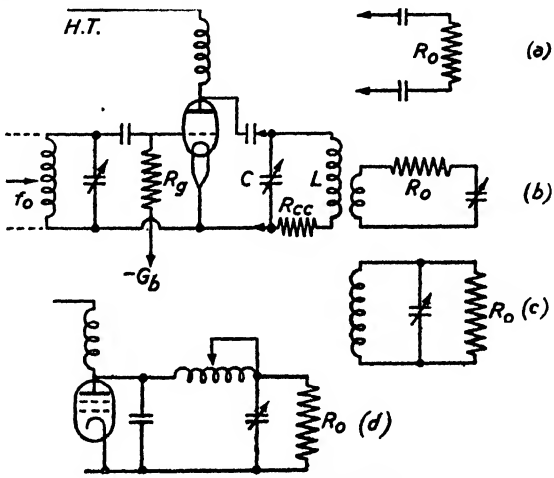


Fig. 208. Amplifier Stage.

through a radio-frequency line. And since the power dealt with in the later stages may be considerable, power efficiency considerations are important.

Fig. 208 shows schematically one stage of a power amplifier, assumed neutralised to prevent self-oscillation, and driven from an input E.M.F., of frequency f_o. The D.C. grid circuit may be completed through a choke, or resistor, and the grid bias obtained from a D.C. supply, or by grid current where a leak is used, or both. Although the circuit is shown choke coupled to the D.C. anode supply, series-fed circuits are quite common. The grid and anode tuned circuits are each tuned to the operating frequency f_o, and the load coupled to the anode circuit either directly or by means of a tuned secondary, the alternative

coupling circuits being shown in "a," "b" and "c," an inductive coupling (at the earthy end) being generally preferred. Sometimes a tapped primary circuit is used and the load connected directly as shown in Fig. 208d, known as a π coupled circuit. Such a system enables a variable coil technique to be employed as no coupling coil is required, but it may be difficult to keep a good balanced circuit (see later) when the load is coupled. We propose to discuss the load circuit first, then the anode circuit as a whole, replacing the valve for the time by an equivalent alternator; then the valve as a converter of D.C. to A.C., and finally the valve cathode and grid circuits.

The Load Circuit

Except in the case of an interstage, where the grid load of the following valve may be regarded as a resistance coupled through a capacitor directly across the anode tuned circuit (coupling "a" above), such a method of coupling the load is rarely used. For the final stage it is usual to employ a resonant secondary circuit with the load either in series, or in parallel, as shown in Fig. 208, "b" and "c." When the load resistance is low, say 100 ohms or less, the series circuit will always be adopted, as it is easy to obtain full loading from the transmitter at unity power factor. If the load is of high resistance, say above 500 ohms, the parallel circuit will normally be found easy to design, as in this case the load resistance is high compared with the reactance of the coupling coil (and of the parallel tuning-capacitance), and full loading can be obtained at a unity power-factor.

For intermediate values of load resistance it is more difficult to decide which type of circuit to use. If the load is connected in series, then, as it increases in resistance, a larger induced E.M.F. is necessary to circulate a given power. If the coupling coil has already been arranged so that the E.M.F. induced in each turn of it is as great as possible, it will be necessary to increase the turns. At the higher frequencies this may result in the coil resonating on its own, due to its self-capacity.

If the parallel connection of load is employed, it can be shown that unless the coil reactance is less than half the value of the load resistance, the value of capacitance which makes the current in the load a maximum will not produce resonance

—that is, the total reactance in the circuit will not be zero. For values of load resistance between about 100 and 500 ohms, therefore, it will be necessary to consider each case on its own merits and decide whether the series or parallel connection will be the most economical.

The Anode Circuit

In Fig. 209 the valve has been replaced by an equivalent alternator, giving a terminal voltage E at a frequency f_c which is the resonance frequency of the anode circuit, the secondary

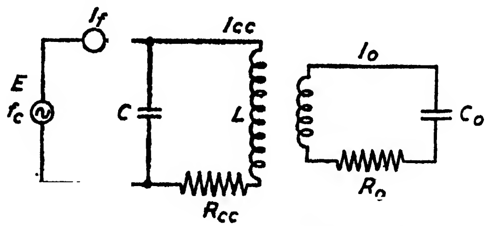


FIG. 209. Equivalent Valve Circuit.

being of the series-tuned type in which R_o represents the useful load, coupled through a mutual M. Which type of circuit is selected for discussion is immaterial, of course, since either form is convertible into the other, as seen from page 183. The circuit $LR_{cc}C$ is usually termed a " tank " circuit, because its principal function is to store energy during those parts of the cycle when the valve may be inoperative, thereby maintaining approximately sinusoidal conditions in LC, even though the current from the valve is far from sinusoidal in character.

Since we have specified that f_c, the frequency of the alternator, is that of the resonance frequencies of the anode circuit tank and secondary, we can simplify the formulæ relating to the circuit relationships thus :

At resonance
$$\omega M I_{cc} = R_o I_o \qquad . \quad . \quad . \quad . \quad (1)$$
$$\omega^2 M^2 I_{cc} = \omega M \, I_o R_o$$
where
$$\omega = 2\pi f_c$$
$$\frac{\omega M I_o}{I_{cc}} = \frac{\omega^2 M^2}{R_o} \qquad . \quad . \quad . \quad . \quad (2)$$

Since the back E.M.F. induced in L, due to I_o, is $\omega M I_o$, then

$\frac{\omega M I_o}{I_{cc}}$ is an equivalent series-resistance in the tank, replacing the effect of the load circuit, the value of this resistance being $\frac{\omega^2 M^2}{R_o}$. Hence the total series resistance R_t in the tank circuit is

$$R_t = R_{cc} + \frac{\omega^2 M^2}{R_o} \quad . \quad . \quad . \quad . \quad (3)$$

We may consider the alternator as being connected across a tank circuit having a series resistance of R_t. Since the tank-circuit condenser will be adjusted so that the A.C. feed current I_f is in phase with E, the whole circuit will be equivalent to a resistance load on the alternator. Thus:

$$Z_e = \frac{\omega^2 L^2}{R_t} = Q_2 \omega L \, . \quad . \quad . \quad . \quad (4)$$

where Q_2 is the apparent Q factor of the tank circuit when the load is coupled.

The currents in the two branches of the tank circuit are practically independent of the circuit losses, even with the lowest Q values in use, and hence we have:

$$I_L = \frac{E}{\omega L} = I_C = \omega C E \quad . \quad . \quad . \quad (5)$$

Since $$\omega^2 = \frac{1}{LC}, \quad I_L = I_C = \frac{E}{\sqrt{\frac{L}{C}}} \quad . \quad . \quad . \quad (6)$$

that is, the current in the tank circuit depends, for any given applied A.C. voltage, entirely on the L/C ratio and not upon the load.

Now $$Q_2 = \frac{\omega L}{R_t} = \frac{\omega L}{R_{cc} + \frac{\omega^2 M^2}{R_o}} \quad . \quad . \quad . \quad (7)$$

and Q_1, the Q factor of the tank circuit alone, is $\frac{\omega L}{R_{cc}}$.

Hence $$\frac{\omega^2 M^2}{R_o} = \omega L \left(\frac{1}{Q_2} - \frac{1}{Q_1} \right) \quad . \quad . \quad . \quad (8)$$

Clearly, our object is to transfer as much of the generator output to the load as possible and it is therefore useful to define

a transfer efficiency as given by $\dfrac{\text{Power in Load, } W_o}{\text{Generator Output, } W_t}$.

The power input to the tank circuit is given by :

$$I_{cc}^2\left(R_{cc} + \frac{\omega^2 M^2}{R_o}\right)$$

of which

$$I_{cc}^2\,\frac{\omega^2 M^2}{R_o}$$

is transferred to the load. The transfer efficiency is, therefore,

$$\frac{\dfrac{\omega^2 M^2}{R_o}}{R_{cc} + \dfrac{\omega^2 M^2}{R_o}} = \frac{\omega L\left(\dfrac{1}{Q_2} - \dfrac{1}{Q_1}\right)}{\dfrac{\omega L}{Q_2}} = \frac{Q_1 - Q_2}{Q_1} \quad . \quad (9)$$

We digress to note that the Q of any circuit, unless Q is very low, can be stated as :

$$\frac{\text{(p.d. across circuit) (current)}}{\text{power}} \text{ or } \frac{V\,A}{W}$$

The p.d. across the circuit is very nearly ωLI and hence

$$\frac{V\,A}{W} = \frac{\omega L\,I\,.\,I}{I^2 R} = \frac{\omega L}{R},$$

which is the voltage magnification factor of the circuit.

From (9) we see that to obtain a high transfer efficiency the tank circuit itself should have as high a Q as possible, but that, when the load circuit is coupled, the effective Q should fall to as low a value as possible. There are, however, a number of design factors to consider and it may be easier to follow if we work through an example, in a somewhat unorthodox way perhaps.

Example (1). A power of 1,000 W, at 10 Mc/s, is required, in a load of 80Ω resistance, from a 3,000 volts R.M.S. supply.

$$I_o^2 \times 80 = 1{,}000\text{W}. \quad \text{Hence } I_o = 3.52\text{A}$$

From (1), $$MI_{cc} = I_o R_o/\omega = \frac{3.52 \times 80}{2\pi \times 10^7}$$

or $$MI_{cc} = 4.5$$

where M is in microhenrys and I_{cc} is in Amperes.

This means we can transfer 1,000 watts into an 80 ohm load, by selecting any convenient value for I_{cc} and using the appropriate mutual. With a given supply voltage, it has been seen from (5) that the value chosen for I_{cc} will automatically fix the L/C ratio of the circuit.

Let us therefore choose various values for I_{cc} and determine the other values which result.

Values of I_{cc} from 1·0 to 20·0 amps. have been selected and the constants of the circuit resulting are shown in Table XII.

TABLE XII. *The Tank Circuit*

	I_{cc} Amps	M μH	MI_{cc}	$\frac{\omega^2 M^2}{R}$ Ω	L $E/\omega L$ μH	C $I_{cc}/\omega E$ $\mu\mu F$
1	1·0	4·5	4·5	1020·0	47·6	5·3
2	2·5	1·82	,,	167·0	19·04	13·25
3	5·0	0·90	,,	40·5	9·52	26·50
4	10·0	·45	,,	10·2	4·76	53·0
5	20·0	·225	,,	2·52	2·38	106·0

In Tables XIII and XIV is shown the resulting values in the circuit, for the " no-load " and " load " condition. In obtaining these we have made only one assumption, that Q_1 remains the same for the different values of coil chosen, a value of 120 being taken as reasonable.

TABLE XIII. *No-load Conditions*

	I_{cc} Amps	R_{cc} Ω	Z_1 Ω	I_f mA	VA	W	$VA/W = Q_1$
1	1·0	25·0	360,000	8·35	3,000	25·0	120
2	2·5	10·0	144,000	20·8	8,500	62·5	,,
3	5·0	5·0	72,000	41·6	15,000	125·0	,,
4	10·0	2·5	36,000	83·5	30,000	250·0	,,
5	20·0	1·25	18,000	167·0	60,000	500·0	,,

TABLE XIV. *Load Conditions*

	I_{cc} Amps	R_t Ω	Z_2 Ω	I_f mA	VA	W	$VA/W = Q_2$	Transf. Efficy.
1	1·0	1045 0	8,800	342	3,000	1,025	2·92	98%
2	2·5	177·0	8,480	354	7,500	1,062	7·06	94%
3	5·0	45·5	8,000	375	15,000	1,125	13·55	89%
4	10·0	12·7	7,200	416	30,000	1,250	24·0	80%
5	20·0	3·75	6,000	500	60,000	1,500	40·0	67%

From these tables the following points should be observed particularly. (*a*) The effective loading varies inversely as the square of the current. (*b*) A fixed goodness of tank circuit, i.e. Q_1 constant, does *not* mean a fixed loss of power in the tank circuit, for a given Alternating Supply voltage, but the tank circuit loss is inversely proportional to the L/C ratio. (*c*) The transfer efficiency is greatest when I_{co} is smallest, but this requires small values of tank circuit capacitance. Also, it is seen from the values of Z_1 and Z_2, that the effective resistance across the generator, when the tank circuit is loaded drops, and therefore a larger A.C. feed current is drawn. Note that the product of voltage and current in the tank circuit remains constant for both loaded and unloaded conditions, if the supply voltage remains constant. In practice, of course, as with any generator, there will be a small progressive drop of voltage as the load is increased.

Choice of Q_2

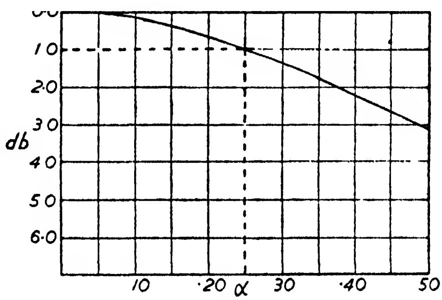
The various factors which bear upon the value of Q_2 selected may be summarised as follows :

1. The lower Q_2, the greater the transfer efficiency.
2. ,, ,, ,, the flatter the frequency response, and the less the side-band cutting.
3. ,, ,, ,, the greater the harmonic content.
4. ,, ,, ,, the greater the divergence between phase, amplitude, and unity P.F. resonance.
5. ,, ,, ,, the greater the L/C ratio.

The first two suggest Q_2 should be lowered, but the other three impose limits, which vary with circumstances. We have already dealt with (1) in a general way, and it will be desirable to discuss the other factors.

(2) *Side-band Cutting.* In designing a tuned-circuit amplifier the receiving engineer has, in addition to obtaining a high voltage amplification, to take two other important factors into consideration, viz. a resonance curve having a pass-band that will accept the wanted frequencies, but having considerable attenuation outside, in order to eliminate interference. Since a transmitter is only producing the necessary band of signal frequencies, we have no need to consider the cut-off region of a

resonance curve, but can concentrate on having negligible attenuation in the pass region. On short waves it is easy to achieve high fidelity as may be seen from Fig. 210 which shows a portion of a universal, resonance curve. It will be observed that f_0 the carrier frequency, f_1 the maximum off-resonance side-band frequency, and Q are related to a factor α such that, for an attenuation of 1 db, α is ·25. Thus at a frequency of 10 Mc/s, even with a Q as high as 100, we could pass side-band frequencies up to 25 kc/s with not more than 1 db loss. In ordinary S.W. transmitters for telephony or broadcasting we



$$\frac{a}{Q} = \frac{f_1}{f_0}$$

where f_0 = resonance frequency
f_1 = max side band off resonance

FIG. 210. Universal Resonance Curve.

need not, therefore, consider (2), but it will become important in facsimile and in television, even on U.S.W.'s.

(3) *Harmonic Content.* When we come to discuss actual valve performance we shall find that, if a valve amplifier is adjusted for high power efficiency, it is not equivalent to a generator giving a sinusoidal output but that the feed-current waveform, may be very distorted, containing large harmonics.

At the resonance frequency the circuit presents a pure resistance given by ωLQ but to the harmonics it presents an impedance which is lower than this—very much lower if the Q is high—the relationship being approximately :

$$\frac{\text{Impedance off resonance}}{\text{Impedance at resonance}} = \frac{1}{Q\frac{f_1}{f_0}\left(1 - \frac{1}{\left(\frac{f_1}{f_0}\right)^2}\right)}$$

where f_1 is the off-resonance frequency and f_0 is the resonance frequency.

In consequence the percentage of harmonics in the voltage wave will be much lower than in the impulsing current wave—in fact, if the Q is fairly high, the output waveform may be practically sinusoidal even though current passes for only a fraction of each cycle. But with low Q values the harmonic output is by no means negligible and would represent a troublesome output on high power transmitters.

For instance, if the current wave consists of a series of half-sine impulses, the amplitude of the second harmonic is $\frac{4}{3\pi}$ that of the fundamental (see page 360). But if we consider the relationship of fundamental and second harmonic, and the impedance relationship given above for the same frequencies, we obtain relative voltages of fundamental E_1 to harmonic E_2 in the output for different circuit Q values, as given below:

Q	50	25	20	15	10	5
E_1/E_2	·0057	·011	·014	·019	·028	·057

(4) *Parallel Resonance.* It is well known that the condition for series current-resonance is $\omega L = \frac{1}{\omega C}$ and is therefore independent of the circuit resistance. With a parallel LC circuit of high Q value its impedance is maximum and resistive at a frequency at which $\omega L = \frac{1}{\omega C}$ (i.e. the series resonance frequency), the generator current is therefore a minimum and in phase with the generator voltage, and there is maximum circulating current in LC, all very desirable conditions and making for ease of adjustment to optimum conditions. (It is such conditions we have assumed in our previous example.) But with circuits of low Q we have more than one so-called resonance frequency; one at which the circuit is resistive, another at which the circuit has maximum impedance, and yet another at which there is maximum circulating current in LC. These various resonances diverge from each other and from the frequency at which $\omega L = \frac{1}{\omega C}$ by an amount which increases with circuit resistance

(i.e. decrease of Q) and is greatest when all the resistance is in one branch, as is usual with a normally-coupled single-valve circuit. Such a condition is clearly undesirable and makes circuits of Q value less than about 10 difficult to adjust. For instance with a circuit of $Q = 10$, the difference between the frequency for series resonance and that for an in-phase current feed is 0·5%. The differences mentioned can be minimised by distributing the load between the two circuit branches and will become least when the resistance is divided equally between them as is possible by circuits which will be discussed later, so that in such cases Q values much less than 10 can be employed if desired.

(5) *Minimum Circuit Capacity.* It has been noted that to get the lowest Q_2 we need the highest L/C ratio, so that our most efficient condition will be when there is no capacity in the circuit additional to that of the valve and "strays." Since circuits usually have to tune to a frequency band, this condition can best be met by using only a very small variable condenser in parallel with the valve and employing a series of appropriate inductors or a variable inductor to cover the wave-range. It is not easy to calculate the minimum stray capacity of a short wave circuit as so much of it is due to random capacity of the circuit as a whole, and not merely the valve capacity.

The design of such a circuit as previously discussed would therefore be carried out in the following manner. An arbitrary value for Q_2 would be selected having regard to such of the above factors as are material, observing that increase of power forces down the value of Q_2, and raising the supply voltage raises Q_2, for reasons which will become evident shortly. As a guide we can state that Q_2 values will lie between 20 and 10 for small powers, and between 10 and 6 for large powers, for a single valve circuit. Having selected Q_2, the value of Q_1 will then be determined to give a transfer efficiency of some 90% with small powers and rather more with large. Then, knowing Q_2 and Q_1, we are in a position to determine the circuit constants.

Thus in the particular example previously set out we should have designed it in the following manner:

***Problem* (2). Deliver 1,000 watts to a load of 80 ohms resistance at a frequency of 10 Mc/s from an A.C. supply of 3,000 volts.**

Solution.

(1) Q_2 = say, 12 $R_o = 80$ Watts output = 1,000
Transfer efficiency
say, 90% $I_o = 3{\cdot}52$ „ tank = 100
$Q_1 = 120$ „ total = 1,100

(2) Find first $VA = Q_2 \times$ Watts
$= 12 \times 1{,}100$
$= 132{,}000.$

(3) Tank current $I_{oc} = VA/E = 132{,}000/3{,}000 = 4{\cdot}4$ amps.

(4) Tank inductance $= E/\omega I_{oc}$. Tank condenser $I_{oc}/E\omega$
$L = 10{\cdot}85\mu$H. $C = 23{\cdot}30\mu\mu$F.

(5) Mutual $= R_o I_o / I_{oc}\,\omega$
$= 1{\cdot}02\mu$H.

(6) A.C. feed current = Watts/voltage
1,100/3,000
·366 amps. } Loaded.

(7) Circuit impedence loaded $= \omega L\, Q_2 = 8{,}100\Omega$.
unloaded $= \omega L\, Q_1 = 81{,}000\Omega$.

It will be observed that the circuit constants obtained lie between those of (2) and (3) of the previous analysis and that the capacitance found ($23{\cdot}3\mu\mu$F) is sufficiently above the estimated value of valve and stray capacitance for a small tuning condenser to be employed.

Raising the Anode Voltage

If we raise the supply voltage, leaving Q_2 the same, the required value of L is proportional to the square of the voltage and C inversely as the square. Thus, in our example, if the voltage becomes 6,000, L would rise to 44μH and C fall to $5{\cdot}8\mu\mu$F. Such a capacitance would be less than that of the valve and circuit self-capacitance, for most layouts, and it would therefore be necessary to raise Q_2.

Power Output

The value of L required is inversely proportional to the power, whilst C is directly proportional to it. Thus if the voltage remained at 3,000 in our example, but the power was to be 100 kW, then L becomes $0{\cdot}108\mu$H and C $2{,}330\mu\mu$F. These figures are clearly unworkable and both a reduction of Q_2 and a rise in voltage will be necessary.

The Valve

We will now deal with the use of a valve for converting D.C. power into A.C. at the correct frequency. Transmitters may employ high or low-impedance triodes, tetrodes or pentodes, but, from the point of view of power efficiency, it is largely immaterial which type is used. This somewhat surprising fact will be made clearer later in the discussion.

Consider the alternator of Fig. 209 replaced by a valve, the circuit being as shown in Fig. 208. Clearly, what we must establish is the relationship between the D.C. anode supply voltage and the A.C. voltage converted from this D.C. by the valve to the circuit and also the waveform of the current supplied by the valve.

Since the circuit *LCR* is resonant and is, therefore, an equivalent resistance load on the valve, we can immediately establish the phase relationships between the currents and voltages, which are as follows:

Anode voltage (E_a) in phase opposition to grid voltage (E_g).
Anode current (I_a) in phase with grid voltage (E_g).
Anode current (I_a) in phase opposition to anode voltage (E_a).
Anode current (I_a) in phase with load voltage (E_{LC}).

These phase relationships will be the same whether the load is a pure resistance or a parallel-resonant circuit, with one important difference. The D.C. voltage drop in the resonant circuit being negligible, the voltage across it is purely alternating and not an alternating voltage superimposed on a D.C. voltage as in the resistance case. Hence the voltage across the valve can rise above the D.C. value during part of the cycle, since the valve voltage is always the difference of supply voltage and anode load voltage. In practice the valve voltage can rise to approximately twice the supply voltage.

In addition to the phase relationships in the valve itself which we have just discussed, we have also those of the oscillatory circuit currents I_L and I_C and voltage E_{LC} which are well known and are those of the ordinary parallel-resonant circuit, namely:

I_L in quadrature (nearly), lagging on E_{LC}
I_C in quadrature, leading on E_{LC}
I_a the " make up " current, in phase with E_{LC}.

Power Output

In order to obtain the greatest output for a given A.C. grid voltage and with sinusoidal anode current and voltage, the load resistance would require to be equal to the A.C. resistance

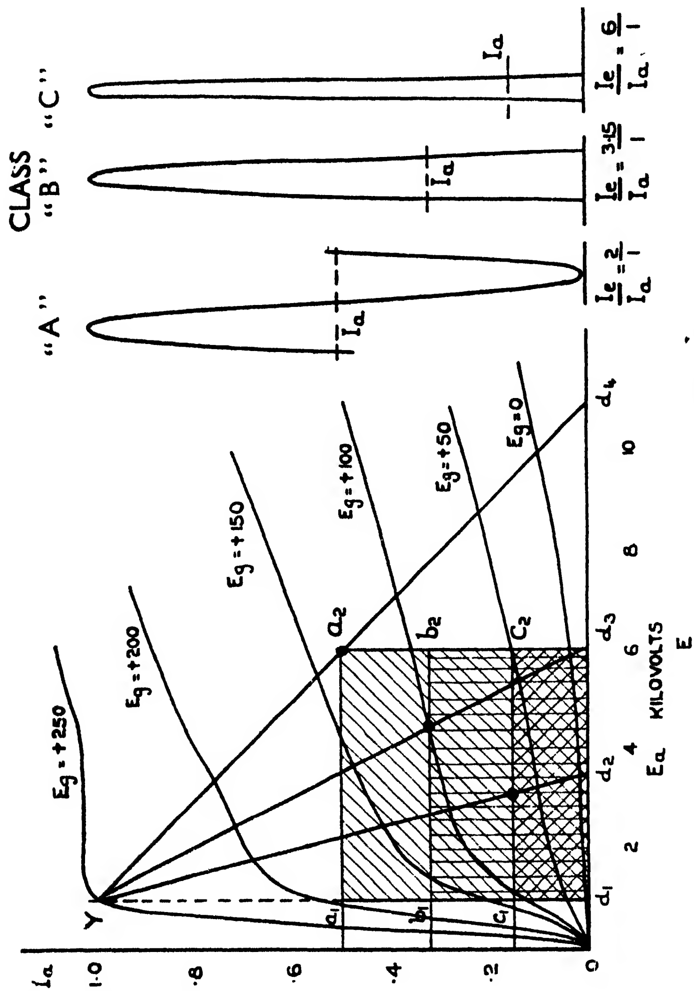


FIG. 211. Types of Valve Operation.

of the valve. This is not usually what we require, however, in power amplifiers, and the load resistance is settled from quite other considerations and is not directly related to the valve resistance.

In practice we cannot afford to avoid the grid current region, as it would reduce our available swing of anode volts and current, and limit the useful characteristics of the valve; and hence the amplitude of A.C. voltage on the grid (usually termed "grid swing") will be made great enough to vary the anode current from zero to the saturation value; further, we never work sinusoidally, but with a current pulse which operates for only a fraction of the cycle to be described later. Under these conditions, the correct load resistance to obtain maximum output is much lower than the valve A.C. resistance, and it is rather a question of utilising as much of the available valve characteristic as possible.

The anode-current/anode-volts characteristic curve of a typical transmitting valve is shown in Fig. 211, where the curve for + 250 volts on the grid represents the "limiting edge" of the characteristic, so-called because curves for larger grid voltages will lie almost on top of this curve for which practically the whole electron emission from the filament reaches either the anode or grid. It will be seen that the available filament emission current in this valve is one ampere.

Class A Valve Operation

We will first discuss the best adjustments on the assumption that anode current is to flow during the whole cycle, that is, the alternating component of the anode current will be approximately sinusoidal, so-called Class A operation.

If the one-ampere, anode current could be obtained when the anode volts were zero, then the operating voltage-current line should terminate at $i_a = 1\text{A}$, $e_a = 0$ and if the D.C. supply voltage is fixed at 6,000 volts then the load line should be a straight line to $i_a = 0$, $e_a = 12{,}000$ volts, so that e_a varied between 0 and 12,000 volts. Such a load would utilise to the fullest extent the available emission and supply voltage. We are assuming that this adjustment is possible without exceeding the permissible anode dissipation.

The actual characteristic shows, however, that 1A cannot be obtained with less than about 1,000 volts on the anode and hence e_a can only usefully vary between 1,000 and 11,000 volts, the amplitude of the alternating voltage being 5,000 peak volts. The best load line is, therefore, Y, a_2, d_4, and in this

case would represent an equivalent resistance $R_e = \frac{5,000}{0{\cdot}5} =$ 10,000 ohms. It will be seen that the mean anode current is 0·5 A and hence the input power is 6,000 × 0·5= 3,000 W and is represented by the rectangle a, a_2, d_3 0. The R.M.S. value of the alternating anode volts is $\frac{5,000}{\sqrt{2}}$ and of the current is $\frac{0{\cdot}5}{\sqrt{2}}$ hence the alternating power output is

$$\tfrac{1}{2} \times 5,000 \times 0{\cdot}5 = 1,250 \text{ watts},$$

represented by half the area a_1 a_2 d_3 d_1 and the efficiency is 42%.

It will be seen that the smaller the minimum anode voltage at which the full anode current can be obtained, the greater is the efficiency, the theoretical maximum being 50% when this voltage is zero. Most transmitting valves are designed so that when the maximum emission is being utilised, the minimum anode voltage will be some 10% to 20% of the D.C. supply voltage, this giving an empirical rule as a basis of design calculations.

The ratio of Peak A.C./D.C. is the voltage conversion of the valve. Although this may be ·85 or ·90 when the valve is delivering peak current, the voltage conversion will rise to nearly 1·0 when the load is reduced, since the peak voltage point slides down the limiting edge from the point Y to 0.

This matching of load to valve (it is rather the reverse in practice, as we have seen) is thus a matter of valve limits and has nothing to do with the valve slope-resistance as can be seen from Fig. 211.

These same arguments apply whether a triode, tetrode or pentode valve is used. The only difference is in the amount of energy required from the grid input-circuit to drive the valve to the peak emission point. In the triode case we have to run into grid current for a considerable portion of the grid cycle, thereby requiring much power from the previous stage. This will be seen by studying the curve of Fig. 211 which is for a triode valve of high slope resistance. In the tetrode or pentode case the zero grid-voltage curve is near the peak-emission

curve, and hence very little grid power will be required to drive the valve to anode saturation.

We can understand easily how the power conversion will fall away with other values of R_e, for if the resistance is lowered we lose voltage swing (steeper load characteristic), whereas if it is raised we lose current swing (flatter load characteristic).

In matching valve and circuit we should aim to have the valve connected across the whole tank, and should not employ "tapped down" circuits. For with a tapped-down circuit the tank current is raised and losses increased; and with circuits mutually coupled to an output, the tap-down increases the harmonic content in output as the impedance to the harmonic frequency is increased in the condenser branch, and reduced in the inductance branch which is coupled to the output.

Since the efficiency of Class A valve operation is low, the method of operating the valve is changed by biassing the grid back negative, so that current flows for only a fraction of the total cycle, the grid driving volts being adjusted so as to drive the valve always up to the peak anode saturation current condition, no matter how much grid negative may be applied. This has the effect of raising the conversion efficiency, and because the valve load circuit is of the parallel circuit type discussed, the voltage across it, and the current circulating within it, remain very nearly sinusoidal even though the feed current supplying it with energy departs very much from a sine-wave form.

As explained, the parallel-resonant circuit acts like a fly-wheel and its efficiency as such will depend upon the ratio between the energy oscillating in it and the energy dissipated in it per cycle, that is, upon the ratio $\frac{\mathrm{k}VA}{\mathrm{k}W}$, or the Q value. The smaller this ratio the greater the efficiency of transfer from tank to output but the greater the harmonic content, as has been shown.

Class B Working

If we bias the grid back to cut-off, in which case the angle of current flow is 180°, this is known as Class B working. The locus of EI will thus be Yd_3d_4, assuming we still drive up to the point Y, this being an equivalent load line,

but not in the accepted sense of the term, namely as a line whose slope is an inverse function of the circuit resistance. It will be seen that as the valve voltage swings down from d_3 to d_1 and back, a half-sine current wave will flow to a peak value Y, as shown under "B," Fig. 211, and that as the anode voltage rises from d_3 to d_4 and down again, no current passes. Thus although anode current and anode voltage are still in anti-phase, the current wave only flows for half the cycle. The equation for series of half-sine waves is :

$$i_a = \frac{I_{max}}{2}\left[\frac{2}{\pi} + \sin\theta - \frac{4}{\pi}\cdot\frac{1}{3}\cos 2\theta \; . \; . \; . \; . \right]$$

The first term is the D.C. component and, when multiplied by the supply voltage, gives the input power. The average current is now $\frac{I_{max}}{\pi}$ = ·318 amps., and the rectangle 0 b_1 b_2 d_3 shows the input power. The alternating voltage is, as we have already explained, very nearly sinusoidal and hence the output power is given by the product of R.M.S. voltage E, by the R.M.S. fundamental current. It will be seen from the series above that the fundamental current is :

$$\frac{I_{max}}{2}\sin\theta, \text{ this having a maximum value} = \frac{I_{max}}{2}.$$

Thus the power output $= \frac{1}{2} \cdot \frac{I_{max}}{2} \mathrm{E}_{A.C.\,max}$.

The remaining terms in the expression for i_a will convey no power because the corresponding harmonics in the voltage wave are negligible. The efficiency is improved to 66% because the anode current now flows mainly when the anode voltage is low, so that the product of anode current and anode voltage integrated over a cycle (the valve anode loss) is smaller.

Class C Working

If the grid bias be still further increased negative, the angle of current flow will now be less than 180°, and we have what is known as Class C valve operation. Thus if the bias is such that the angle of current flow is about 120°, the *EI* line will be approximately $Yd_2d_3d_4$, Fig. 211, a current pulse of shape indicated by "C," Fig. 211 passing only as the anode voltage

falls from d_2 to d_1 and back to d_2, current being cut off during the rest of the cycle of anode voltage up to d_4 and back to d_2. Thus the input power has fallen to that represented by the rectangle $0\,c_1c_2d_3$, the conversion efficiency has risen above that for Class B working, but actually the output power would have fallen somewhat with the same peak current Y shown in Fig. 211.

Quite clearly with the three types of adjustment we have widely varying conditions, not only of efficiency but of valve current conditions and input power, Table XV below giving some results in tabulated form.

TABLE XV. *D.C. Supply Voltage* 6,000 V. *Peak Emission* 1A. *Minimum Anode Voltage* 1,000 V. *Alternating voltage* (*max. value*) 5,000 V.

Class	Anode Current.		Ratio Peak/ Mean.	Imput Power Watts.	Anode Loss.	Output Power Watts.	Efficiency %
	D.C. Compt.	Fundamental A.C. Compt. (max. value.)					
A	0·5	0·5	2	3,000	1,750	1,250	42
B	0·318	0·5	3·15	1,890	640	1,250	66
C (½ sine wave.)	0·163	0·310	6·15	978	203	775	79

Design of Circuits for Class B and C

It is clear that for any type of valve operation other than Class A, the power output on a frequency f_o depends upon the A.C. component of the fundamental frequency f_o in the current wave, the relationship of this component to the peak and average valve currents, upon the resulting power conversion efficiency, and upon the valve characteristics, which may be considered to be very nearly linear when operated for a portion of the cycle up to saturation.

As has been seen from the worked-out Class B example, the various factors are directly a function of the wave-form, and since we consider the valve linear and grid and anode voltage swings sinusoidal, they are directly a function of the "angle of current flow." By angle of current flow is meant

that portion about the peak of a sine wave whose base is $\theta°$. Thus with Class B, $\theta°$ is 180° and we have a complete half-sine wave with its peak at 90°. With a Class C, 120° angle of current flow, the base is 120°. The shape of the pulse, therefore, is that portion of a sine curve from its peak at 90°, to 30° one side and 150° the other.

The relationships wanted can best be shown by a series of curves, Fig. 212 showing the conversion efficiency plotted as a function of angle of current flow, $\theta°$, and Fig. 213 shows the cur-

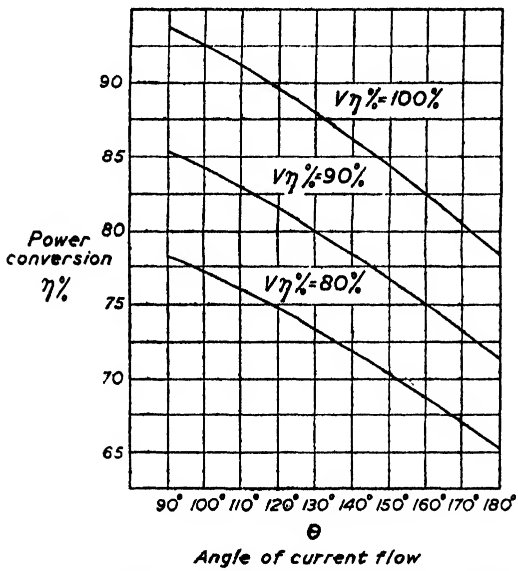


FIG. 212. Power Conversion for Different Angles of Current Flow.

rent relationships also plotted against an angle of current flow $\theta°$.

Three efficiency curves are given, one for each of three voltage conversions, of 100%, 90% and 80%, so that the peak and R.M.S. A.C. voltage can be directly related to the D.C. supply voltage available.

The current curves are related to a given value of A.C. fundamental current, of 1·0 A peak value, not to be confused with the peak saturation current. There are a number of interesting points to call attention to in Fig. 213. First, that for a given value of alternating current, the peak current required for Class B working is exactly the same as required

for Class A working, as may be seen by comparing peak points at 180° and 360°. That for angles of current flow between 360° and 180°, a less peak is required to give the same A.C., whereas for angles of current flow below 180°, the peak current required rises very rapidly. Observe that the average current

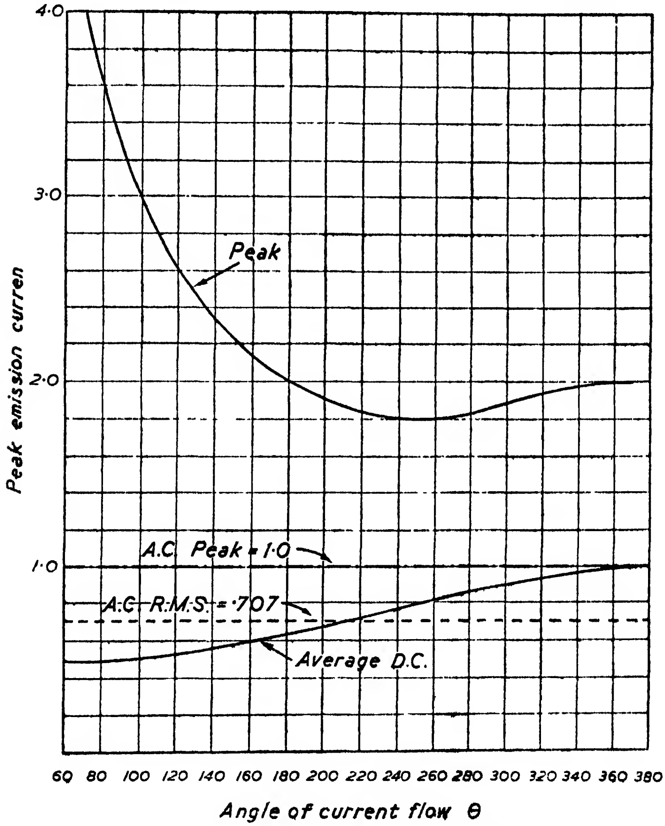


FIG. 213. Relation of Peak Currents for Different Angles of Current Flow.

value, i.e. the effective D.C. value, falls steadily as the angle of current flow is reduced, and that at an angle of current flow of 210°, the D.C. component equals the R.M.S. A.C. component. For angles less than 210°, the D.C. value is actually less than the R.M.S. A.C. being produced.

Once we have settled upon the angle of current flow, by the use of the curves given in Figs. 212, 213, it is a simple matter to

determine the valve parameters required, and it will be easiest to show how they can be used by continuing the design of the transmitter, the output circuit of which was given on page 354.

Problem 3. Determine the valve parameters and supply, for a valve Class C operated, $\theta = 130°$. The output and tank circuits have characteristics as previously determined, viz. :

Frequency $f_o = 10$ Mc/s. $R_o = 80\Omega$. $I_o = 3{\cdot}52$A.
Tank : $L = 10{\cdot}85\ \mu$H. $C = 23{\cdot}30\ \mu\mu$F. $I_{cc} = 4{\cdot}40$A. $Z = 8{,}100\Omega$ loaded.
Power : Output 1,000 W.
Tank = 100 W.

	Total 1,100 watts	(1)
	A.C. Voltage (R.M.S.) = 3,000	(2)
	A.C. Current (R.M.S.) = E/Z = ·366A.	(3)

Solution 3. The angle of current flow $\theta = 130°$. Assuming a reasonable voltage conversion of ·85, then from Fig. 212, interpolating between curves of 80% and 90%, we find that for $\theta° = 130°$, the conversion efficiency = 76·5%.

Power

Total A.C. power required	W_o	= 1,100 watts.
,, power to be supplied, 1100/·765	W_i	= 1,440 ,,
Valve loss	W_a	= 340 ,,

Voltage

Voltage conversion (as above)	= 0·85
A.C. R.M.S. volts required	= 3,000 volts.
A.C. peak volts required, 3,000/·707	= 4,243 ,,
D.C. supply voltage, 4243/·85	= 5,000 ,,

Current. From the curves of Fig. 213 it is seen that for an angle of 130° we have the following current relationships :

Peak saturation	2·46
A.C. current, peak	1·0
A.C. current, R.M.S.	·707
D.C. average feed	·560

Since we require an A.C. feed of ·366A, we must reduce the above current figures in proportion and thus we obtain :

A.C.	A.C. feed, R.M.S., ·707/1·935	= ·366
	A.C. peak, 1·00/1·935	= ·517
D.C.	Peak saturation, 2·46/1·935	= 1·27
	Average current, ·56/1·935	= ·289

Thus we have to select a valve having an anode dissipation of 340 watts, with a cathode capable of giving a peak anode current of

1·27 amps., that is more than four times the average value, at an anode voltage of 750, i.e. 15% of the D.C. voltage. Observe that the impedance and type of valve do not enter into the problem at this stage. The anode D.C. supply needed will be 5,000 volts.

The choice of θ is determined chiefly by the requirements of the transmitter although the power and choice of valves available also enter into the matter. For a telegraph transmitter, where we are only concerned with "on" "off" conditions, we work always with Class C, and the value of θ chosen will lie between 120° and 150°, the former giving the higher anode conversion efficiency, as has been seen from the curves of Fig. 213. But from the curves of Fig. 214 it is also seen that as θ is reduced below angles of 150°, the peak anode current rises very rapidly, and hence greater filament power will be required. Further, the smaller θ, the more the grid negative required, and in consequence the greater must be the grid driving voltage and power. Hence, although the smaller the value of θ the greater the anode conversion efficiency, but not necessarily the overall efficiency, and it has been found in practice that 120° represents a normal minimum for θ.

Load Impedance and Load Line

We have seen that the parallel LCR circuit acts as a resistance load, and it is well known that such resistance can be drawn across the E_aI_a valve characteristic, as a line of slope E/I (reversed). This was done for the Class A example in Fig. 211. When we bias the grid negative so that current flows only for a portion of the cycle, although we may be supplying the same fundamental frequency power to the same circuit at the same average voltage, the *circuit* slope line no longer shows the operating condition, or traces out the changes of E and I, because the locus of the EI changes only coincide with the impedance slope line under sinusoidal conditions. How the various EI lines change with grid bias can be seen from Fig. 214 which are plotted for a fixed D.C. voltage, constant circuit impedance and the same A.C. output. The curves show that the active part of the EI curve (i.e. that part from saturation to cut-off) commence by flattening slightly, then steepen, and at $\theta = 180°$ the curve is twice as steep as the load line, since it has the same peak value of current and half the

voltage base. And for smaller values of θ than 180°, the curves steepen rapidly, becoming nearly vertical for very small values of θ.

It so happens that for $\theta = 180°$ the *EI* curve is twice as

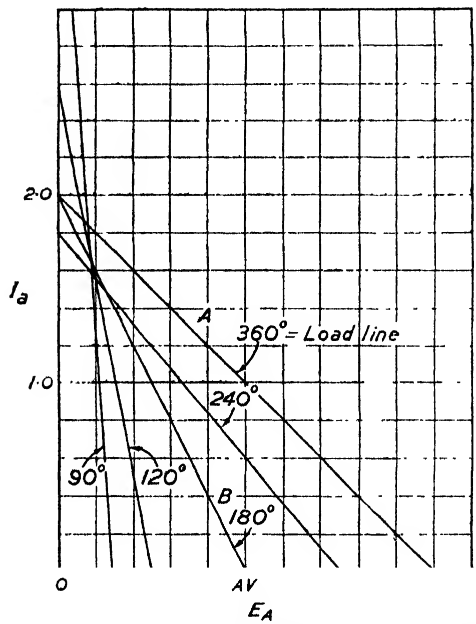


FIG. 214. Operating Conditions for Different Angles of Current Flow.

steep as the load curves, and there is therefore a direct relationship between them and the angles, viz. 360/180. But for other values of θ there is no simple relationship.

Off Resonance Conditions

Before leaving the subject of the output circuit we must discuss the effect of not correctly tuning the output circuit to the input frequency f_o (which remains constant as it is determined by the driving source).

The impedance/frequency and phase/frequency curves of a parallel resonant circuit are dependent upon the Q of the circuit. This means that for an unloaded tank, where the Q is high, we shall have a sharp-peaked impedance curve, as shown by Q_1, Fig. 215, but for a loaded tank its shape will be

much flatter, as shown by Q_2, Z_1 and Z_2 being the effective resistance values at resonance, when the phase angle is zero.

Since the A.C. feed $I_f = E/Z$, the shape of the feed current curves would be as shown by I_1 and I_2, if the circuit was supplied from a constant voltage source. These curves show that the feed is a minimum at resonance, but that with a lightly loaded circuit there is a very considerable dip in the feed curve, whereas with a heavily loaded circuit the dip may be small.

These curves of feed will in the main be similar to those of the D.C. feed in a Class C operated valve, not quantitatively, however, since with a valve conversion the A.C. voltage

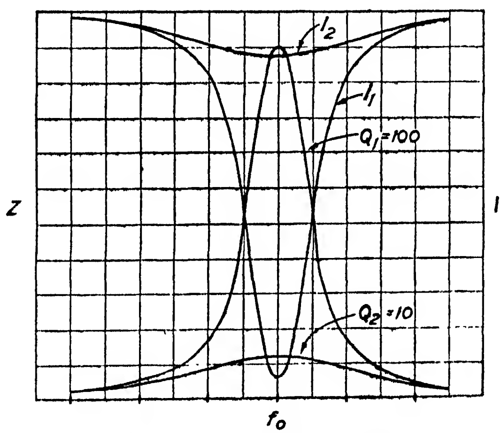


FIG. 215. Current-Frequency Curves for Tuned Output Circuits.

is not constant under all conditions. In fact as the load circuit becomes reactive and of lower impedance, off tune, the *EI* locus becomes elliptical in form and its axis steeper, which means that not only do *E* and *I* change from their anti-phase condition, but the amount of D.C. voltage converted is greatly reduced.

We have mentioned the importance of the opposite-phase condition with Class C and, as it is not easy to pick out the minimum-feed condition with a heavily loaded circuit, it is preferable to tune the circuits up at less than full coupling, where the tuning is clearly observed. In tuning a coupled-circuit transmitter it is necessary to ensure that both circuits are correctly tuned. Thus the primary should be tuned

without a secondary (if necessary on reduced power), and the secondary tuned afterwards. Otherwise if the two are coupled first without tuning, there are a number of compromise tuning adjustments which will apparently give a resonance adjustment, but which in reality are incorrect.

The Grid Input Circuit

The grid of an amplifier requires both a steady negative bias and an H.F. potential, correctly proportioned so as to obtain the correct angle of current flow and the required power output. Actually there is a fair latitude in grid-circuit adjustment, as will be seen later.

When using Class B for telephony the static bias must be obtained from a D.C. supply, as any automatic grid-current bias would be subject to unwanted variations when the input is modulated. For Class C telegraph operation, however, although a small static bias may be provided to safeguard the valve should the H.F. driving voltage fail, it is usual to obtain most of the bias from a grid leak. This resistance carries the pulses of grid current which occur when the grid goes positive each cycle and the average value of these pulses through the bias resistor provides a bias voltage.

The H.F. voltage is generally obtained from the tuned circuit of the previous stage, although in some cases the output of the previous valve is coupled to the following tuned grid circuit through a H.F. transmission line. Quite clearly the applied grid voltage must be of sufficient value for its positive peak (however loaded by grid current) to drive the grid up to the limiting edge. As previously mentioned, the smaller the angle of current flow the greater must be the grid negative bias and the greater the H.F. grid voltage.

It was mentioned earlier that since we must run into grid current in order to obtain a reasonable anode conversion efficiency, the driving stage is called upon to provide H.F. power of an amount which depends upon the type of valve, the frequency of operation and the way it is operated.

If we had the full characteristics of the valve, including the values of grid current at high, positive, grid-voltages and low anode volts, it would be possible to design the grid circuit

directly, but as these valve constants are never available values are usually obtained empirically.

If no characteristics are available a rough approximation can be made in the following way.

Fig. 216 shows the conditions obtaining at the minimum anode-voltage condition in an amplifier in tune. E is the D.C. anode-voltage, E_{ac} the half anode-voltage swing falling to a value of E_a min., taken as 15% of E. i_a is the anode-current pulse, I_{dc} being its average. E_{gb} is the grid

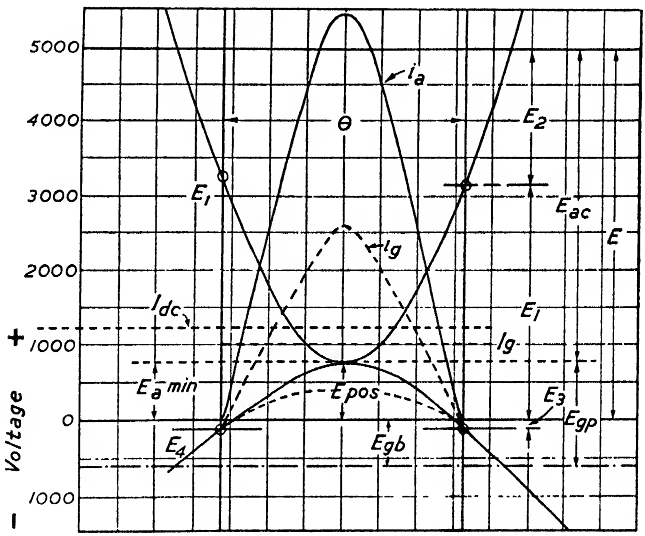


FIG. 216. Valve Operating Conditions.

negative bias, E_{gp} the grid swing (sinusoidal about the bias datum), and E_{pos} the positive grid swing above zero (assumed sinusoidal), is the grid-current pulse and I_g the average grid current. θ is the angle of current flow and E_1 is the value of anode voltage at anode current cut-off.

It will be observed that the anode and grid swings have been assumed sinusoidal and that the positive peak value rises to that of the minimum anode voltage. This is a very approximate empirical rule since in actual fact the positive grid voltage peak is considerably flattened, due to grid current.

Thus, if peak grid measurements taken positive and negative from zero are added, they will not be found to equal double the grid bias, but less due to this flattening.

To obtain these curves we need to know only the μ_a of the valve and the angle of current flow θ. Thus, considering first the anode circuit, we have:

$$E_2 = E_{ac} \cos \frac{\theta}{2}$$

$$E_1 = E - E_2$$

$$E_1 = E - (E - E_{min}) \cos \frac{\theta}{2} \quad . \quad . \quad (10)$$

E_1E_1 is the anode voltage at the current cut-off points and hence we require at these points grid voltage negative values of:

$$E_3 = E_4 = \frac{E_1}{\mu_a} \quad . \quad . \quad . \quad . \quad (11)$$

Since we are assuming that $E_{pos} = E_{min}$, we have three points lying on a cosine curve and can evaluate these points from:

$$E_{gp} = \frac{E_{pos} - E_3}{1 - \cos \frac{\theta}{2}} \quad . \quad . \quad . \quad (12)$$

and $$E_{gb} = E_{gp} - E_{pos} \text{ approximately} \quad . \quad . \quad (13)$$

Grid Bias

As mentioned previously in Class C amplifiers the grid negative bias is obtained from the IR drop down the grid resistance. It might be thought, therefore, that the average grid current value when driving to the limiting edge would be a fairly definite value and that this would determine the grid resistance value rigidly. But it is found in practice that quite a fair range of resistance values are possible and that the grid current value can be varied to give the required IR value without a great deal of change occurring.

Actual best values depend to some extent on the valve size, the larger the valve the smaller the resistance. With small and medium size valves resistances between 30,000 and 15,000Ω

are usual, but with the larger water-cooled valves the value may be as low as 4,000Ω. The higher the resistance used the smaller the losses but the more the tendency for the amplifier to squegger (see page 386).

Grid losses, apart from the tuned circuit, are of two types, the valve loss and the grid resistor loss. There are losses in the valve due to the fact that the grid current pulse is in phase with the grid positive voltage peak, and their integrated product results in heating loss at the grid. This loss will be proportional to the square of the grid current. The loss in the grid resistance is due to I^2R losses, and here again the loss is proportional to the square of the current. Since for a given grid bias negative, I_gR_g is constant, and as I_g is therefore inversely proportional to R_g, increase of R_g keeps down the loss.

Overall grid losses are usually of the order of 5% to 10% of the anode power output up to the limit of frequency for which the valve is designed. If operated at higher frequences the grid losses rise rapidly, for various reasons which are discussed later in the chapter, and can be even of the same order as the anode output.

Input Voltage and Amplifier Linearity

It will have been realised that to drive a valve up to the peak condition that we have specified as most efficient for power conversion requires a given input voltage. Since also we have to drive the valve through the grid current region the driving stage will have to provide power of an amount which is determined by the type of valve used and the requirements of the circuit. The relationship of input voltage to output current and amplifier efficiency is shown in Fig. 217, which show clearly that as the driving voltage is raised both output current and efficiency rise from zero to a saturation value, beyond which there is no great change except in the value of grid current and input power, which continue to rise. These curves, curiously enough, are fairly similar whether the amplifier stage is biassed by static or grid-leak bias except that with the leak bias the input power rises to greater values since, as the drive is increased, the current in the grid forces the grid to greater negative values and much more power is necessary to over-

drive the amplifier in consequence. And there is a less linear relationship between input volts and output current. From these curves it is clear that except for increasing input power to the grid, an amplifier can be overdriven without much difference being observed, the increased power merely being absorbed in grid loss, whereas underdriving leads to a reduction of amplifier output and efficiency. With a telegraph transmitter it is essential to operate always at the full peak

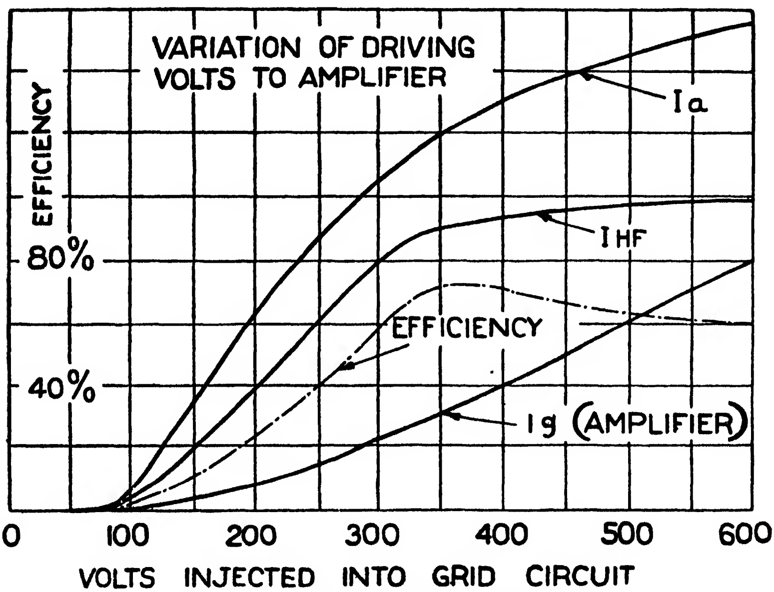


FIG. 217. Effect of Varying Input Voltage.

condition on " mark " so as to obtain the maximum efficiency of the amplifier, the input being adjusted to give just that condition so that the greatest power stage gain can be obtained. Since we are only concerned with " mark " and " space " conditions, the shape of the input-output curves is of no importance and in general they will not be linear with the usual Class C-operated valve.

We may, however, be concerned with grid-modulated carriers and in this case the linearity of the amplifier will become of importance. Methods for obtaining linear conditions will be treated in the next chapter.

The Cathode Circuit

Although oxide-coated filaments are used on low voltages, they have not generally been found successful above some 3,000 volts, except in the special case of pulse-operated valves, and in consequence the great majority of transmitting valves employ a tungsten filament. This is because the oxide-coated cathode is susceptible to poisoning of the surface due to returning ions, which tungsten is not, and although all transmitting valves are of the hard type, the higher the operating voltages the more difficult it is to free the valve from ionisation. Ionisation is inversely a function of gas pressure but directly a function of electron velocity. But electron velocity is proportional to the square of the operating voltage, so that if we double the anode voltage we would need to increase the goodness of vacuum four times for the valve to remain in an equivalent condition as far as ionisation is concerned.

Ionisation itself may not be harmful, as witness the gas-filled rectifier with oxide-coated filament, but here there is no potential to speak of across the valve, and in consequence the returning positive ions have insufficient velocity to damage the cathode, but it is the high velocity ions in the low pressure, hard valve, which do the damage.

A second failing is due to contamination of the grid with active material resulting in primary grid emission at a comparatively low temperature.

The fact that tungsten filaments are used means that the filament power required is not inconsiderable, and a good working figure is 5% to 10% of the power output of the valve. The cathode filament used will always be of the low-voltage high-current type. This is desirable, not only because it then approaches an equi-potential cathode, but the high current is necessary because Class B or C operation demands high peak emission. Thus, except for small powers, quite a large filament is necessary, and special precautions have to be taken in operating these large filaments if a good life is required.

The resistance of cold tungsten is only about 1/12th that of its value when operating at the normal running temperature of about 2,500° K. Thus, if normal filament volts are switched on, a high surge of current results (some 1,400% above normal), this current surge falling to a normal value in a time which is

in direct proportion to the diameter of the filament wire. Thus with wires carrying a normally heavy current, the surge will be so large and persist so long that the strong magnetic field set up will strain and probably break the wire. This is certainly so in the large water-cooled valves, and the effect is sufficiently serious with filament diameters down to ·65 mm.

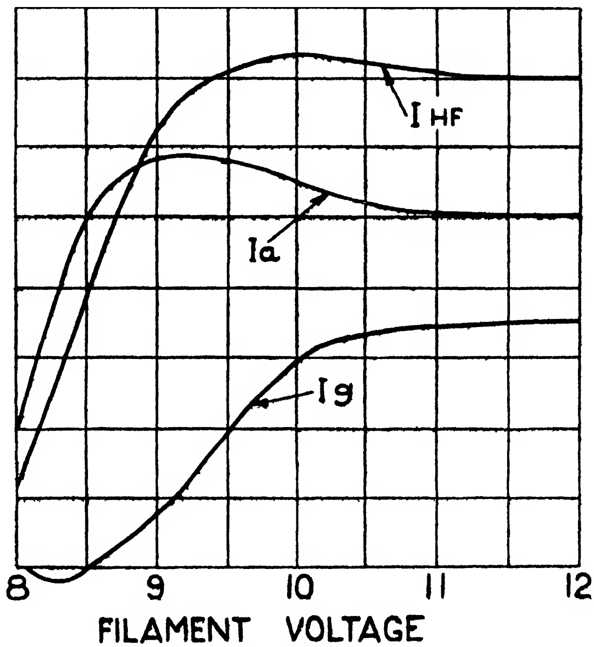


FIG. 218. Effect of Varying Filament Voltage.

to warrant precautionary measures being adopted, if long life is desired. These precautions consist in switching on the filament voltage in stages, the amount of resistance and the number of stages being so arranged that at no time does the filament current exceed some 15% above normal. With small filaments, since the current surge falls rapidly, this can be accomplished in one step, but with the larger filaments two or three stages, each properly designed with the correct time-lag, become necessary.

Filament Emission

A transmitting valve is very sensitive to filament emission limitation, and if this emission is reduced below a certain value no output at all can be obtained, the curves of Fig. 218 showing the effects resulting from reduction of filament voltage.

It is observed that below a certain critical value, indicated by a rapid fall of grid current, the output falls away rapidly, and this is due to secondary emission taking place in the valve.

It is a well-known effect on long waves and appears to be much more marked on short waves, insufficient filament emission being the cause of many transmitters not functioning efficiently, particularly when a valve is ageing.

Push-Pull Circuits

The name " push-pull " is given to those circuits employing two valves arranged differentially. Of course it is not necessary to employ two valves to obtain a push-pull action, nor does

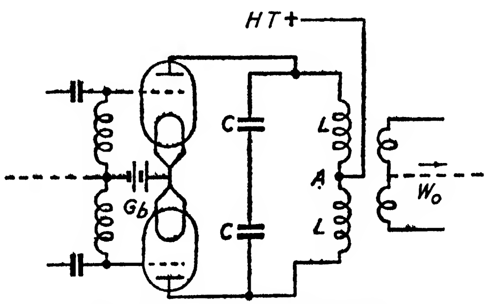


Fig. 219. A Push-Pull Circuit.

the name adequately describe the circuit to which it is applied, for it is possible, with small modifications of a two-valve circuit, to produce a variety of effects, some of which are not push-pull in character.

We will assume the reader to be familiar with the general principles of such push-pull circuits and merely give a brief introduction before discussing the application of the circuit to the short-wave amplifier.

Fig. 219 shows a schematic push-pull circuit, the input and output circuits being centre-tapped as shown. Since the source of H.T. is connected to *A*, the centre point of the output

circuit, the main feed divides at this point, and the feeds to each valve flow in opposite directions through the output circuit. Thus as long as these feeds remain equal or change equally no A.C. output can result. But any inequality of feed to the valves results in output, and a fall of current through one valve directly assists a rise of current through the other, since the feed currents are flowing in opposite directions through the output.

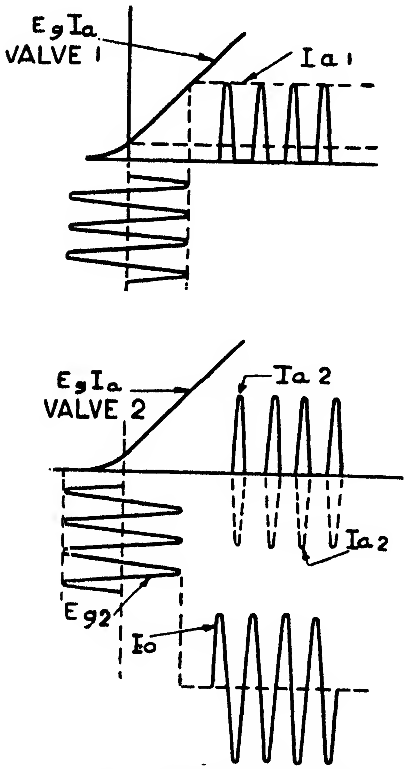


FIG. 220. Illustrating Push-Pull Operation.

Turning to the input side and considering the application of an H.F. voltage, it will be seen that the voltage to each valve can be but half the total input voltage, and hence, from a given available E.M.F., each valve of a push-pull combination can get but half the input volts a single valve could get. Any input acts differently on the grids of the two valves, however, the voltage on grid 1 rising less negative as that on grid 2 rises more negative, and vice versa. This means that the feed to valve 1 will flow in opposite phase to the feed to valve 2, and, as shown in Fig. 220, such effect will be additive in the output circuit.

If the valves are each biassed to the bottom of the linear part of the characteristic (as shown) so as to eliminate the D.C. feed (or nearly so), which is the usual adjustment, we get a push-pull action from the valves; for if No. 1 valve "pushes" for the one half cycle, when its grid goes less negative No. 2 valve "pulls" for the alternative half cycle. The fact

that there is partial or complete rectification of the feed to each valve will make no difference to the push-pull action.

It is, of course, essential that each valve shall have the same constants, and if a linear amplifier is required we cannot bias the valves to the cut-off point, but only above the bottom bend, so that the total characteristic is linear. If the swing on each grid is limited to a value which avoids the grid current region, such an amplifier is often called Class AB, because its overall performance comes under Class A, but each valve is operating under Class B conditions.

If the valves are set back beyond the cut-off point, so that each is operating in a Class C manner, the feed no longer follows the input wave-shape, but because the differential action of the valves and circuits produce a symmetrical wave-shape at all times, no even harmonics are present. Thus Fig. 221 shows the output wave resulting from a "push-pull" circuit biassed beyond cut-off, and supplied from a sinusoidal input voltage. Actually, an output of even harmonics can be produced if there is any "carry through" the valve, due to a voltage appearing between the centre point of the output and earth. It is for this reason that screen-grid valves are to be preferred to triodes for frequency-doubling, which will be discussed later.

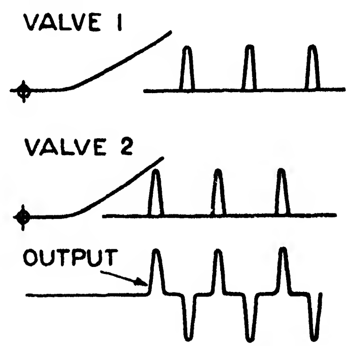


FIG. 221. Illustrating Class C Push-Pull Operation.

Push-Pull Power Amplifier

To indicate the main differences caused by the adoption of push-pull, it is easiest to consider the modifications in working made to the amplifier discussed earlier in the chapter if we add a second valve in push-pull.

Since the D.C. anode voltage is now common to the two valves, shown fed to the centre point of the output, the circuits associated with each valve are in one sense in series. Thus if we assume each valve associated with the same L.C. values as

before determined, i.e. the capacity is nearly at the minimum value possible, the closed circuit current is the same value as are the volt-amperes for each, but since the load is now shared between the valves, each taking half, the ratio VA/watts (or Q), for each, is doubled. Looking at it another way—since the A.C. volts across the whole circuit are doubled (because the total L is doubled and total C halved), whilst the watts remain the same, the Q_2 of the whole circuit is doubled.

It is seen that if the capacitance associated with each is at a minimum, we should not be able to reduce the Q_2 to the same low value as with the single valve. This, however, is not a fair comparison because each valve is only called upon to deliver half the power, and we could therefore use smaller valves with less inter-electrode capacitance. We have effectively doubled the A.C. voltage by the adoption of this push-pull circuit. Thus, although with low-power circuits we may find it necessary to increase Q_2, with medium and high powers there is no difficulty in reducing the value of Q_2 to any low value desired. In fact with large powers the push-pull circuit is essential to enable us to maintain a reasonable inductance value and to increase effectively the working voltage. On the assumption, then, that we can choose as low a value of Q_2 as we like, let us discuss the factors which limited its reduction with the single-valve circuit, namely, harmonic content and divergence of the parallel resonances.

As regards harmonic content, we have seen that a correctly adjusted push-pull circuit eliminates all even harmonics and hence the second, which is by far the most troublesome, disappears. Thus with the lowest Q_2 value no second harmonic will appear in the output, and even a Class C-operated valve may be regarded as acting almost as a sinusoidal generator. It is true that the third harmonic is still present, but from the formula on page 52, and making an analysis of the wave-shape shown in Fig. 220, with angle of current flow of 120°, we find the ratio of third harmonic to fundamental is rather less than ·01 for a Q_2 value of 10.

Considering the parallel resonances, it was pointed out that the divergence of these resonance frequencies from one another was reduced to a minimum when the circuit resistance was evenly divided between the two branches. In the push-pull

circuit, because of the differential action, we have a centre-tapped circuit and do therefore reduce this effect to negligible proportions.

Reviewing the subject generally, although all the remarks made earlier in the chapter apply, we see that the limitations of Q_2 reduction do not apply nearly to the same extent, and Q_2 values can, if necessary, be reduced to values as low as 5 or less. This does not mean that we shall use such a low value, and with medium and low-power circuits the value chosen will still be in the region of 10.

Otherwise, as regards cathode loss and input, the push-pull circuit is similar to a single-valve circuit, when compared on a basis of a given power output. For if we use two valves in push-pull we should only require half the emission from each valve, and a correspondingly smaller filament in each, and since each valve is handling less power the input voltage will also be reduced.

A very important additional advantage of the push-pull circuit lies in the symmetry of the circuit that can be designed and this point will now be dealt with.

Neutralisation Circuits

It has already been mentioned that an essential part of a power amplifier system is the provision of methods for preventing self-oscillation. As in receiving work these consist in the use of adequate screening, the careful layout of components and wiring and the choking and by-passing of common supply circuits. In the case of triode amplifiers we have, in addition, the neutralisation of the grid-anode, feed-back capacity.

Triode Anti-Reaction Circuits

The simplest arrangement of triode neutralisation is as shown in Fig. 222a (diagrammatically in Fig. 222b). Here the input coil is split and the filament tapping taken to the centre so that grid/filament is across but half the input coil L_g. One end of the coil is already coupled to the output circuit through the grid/anode capacity of the valve C_{ag}, and hence if the other end of the coil is connected through the condenser C_n to the anode it will act as an anti-reaction condenser, and balance the grid/anode capacity coupling. Thus power in the output

changes the potential of the anode relative to the filament and hence will create currents through the condensers C_{ag} and C_n which flow to earth through the input coil in opposite phase,

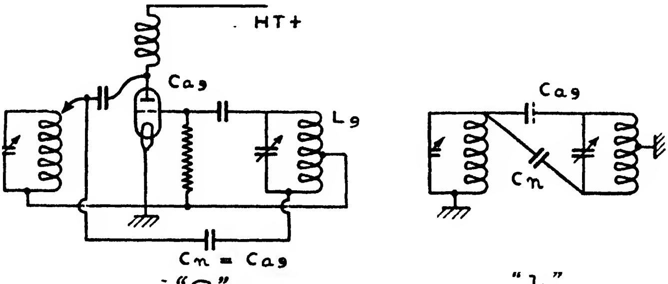


FIG. 222. Neutralisation by Centre-tapped Grid Circuit.

and so cause no feed-back voltage between grid and filament.

An alternative arrangement is shown in Fig. 223a (diagrammatically in Fig. 223b). In this case the whole input is left across grid-filament, but the output centre-tapped instead,

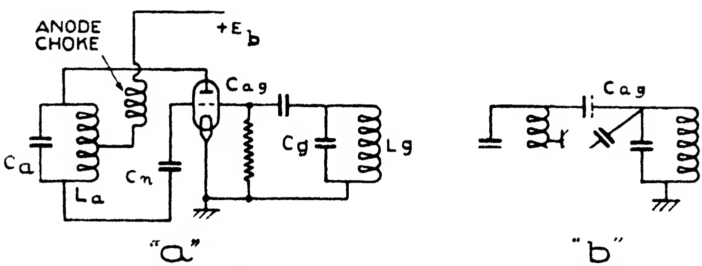


FIG. 223. Neutralisation by Centre-tapped Anode Circuit.

and the free end connected back to the grid. Since the anode supply will be connected between filament and the tapping point on the coil, the circuit is now at high potential, and this is undesirable on high voltage systems. Otherwise the action is similar to the previous arrangement. It is a matter of convenience which method is adopted and both are found in short wave circuit practice, but on very short waves and large powers this simple balanced arrangement is not too satisfactory.

Examination of the above circuits shows that although the

valve capacity has been neutralised, neither circuit is symmetrical as regards earth and any asymmetry of circuit is undesirable. In fact, ordinary methods of balancing are of doubtful value if high efficiency is required, because valve capacity is not the sole factor to contend with, circuit layout is more important.

If we build a geometrically symmetrical circuit, with earth as a datum, we have the best chance of obtaining stable results on very short waves, provided all connecting leads are reduced to a minimum; which is another way of stating that in balancing a circuit to achieve zero feed-back, not only must all obvious coupling be balanced but capacity effects to earth must not be forgotten. As our minds are not accustomed to think in terms of the minute values which are only significant at these very high frequencies, the mechanically balanced layout automatically helps us to obtain the desired results.

The Bridge Balance

C. S. Franklin developed a truly balanced system by building a circuit from a bridge point of view, that is to say, taking the valve capacity C_{ag} as one arm of a bridge, three other capacities are provided to complete the bridge and the whole circuit built

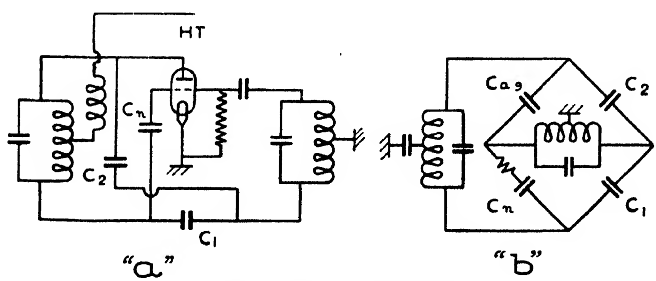


FIG. 224. Simple Bridge Balance.

to fit this symmetrically, not only as regards input and output but with due regard to the reference point earth, so that no coupling exists between circuits placed across the diagonals of the bridge arms, as shown in Figs. 224a and 224b. On wavelengths not too short and small powers, the circuit as shown will give a clean balance; but on waves below some 25 metres,

the circuit in the simple form described above is not perfect, owing to the fact that the bridge shown is but a pure capacity balance and no account has been taken either of the power factor of the condensers making it up or the resistance of the conductors. Since the valve is a leaky condenser, it is necessary to add resistance to the complementary bridge arm to compensate for this, and either a large shunt resistance or a small series resistance can be used. It is simpler to use a series resistance, as only a fraction of an ohm is required, and with this added compensation a perfect bridge balance can be obtained.

It is but a step to turn the single valve bridge circuit to a two-valve bridge for the purpose of handling more power, the

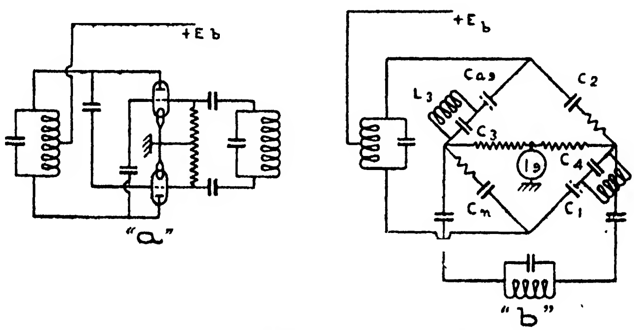


FIG. 225. Modified Bridge Balance.

condenser C_1 being replaced by a second valve, as shown in Figs. 225a, 225b, and this circuit will now be recognised as a push-pull type, symmetrically balanced for anti-reaction.

The layout of the bridge circuit must be such that leads to the grids of the valves, and in fact leads generally, are made as short as possible. On very short wavelengths, however, the length of lead into the valve may be sufficient to upset the balance.

For although the corners of the bridge arms are in opposite phase the reactance of the leads from the corners to the valves throws out the anti-phase voltage condition on the grids. To compensate for this, condensers are necessary in each lead whose reactance will cancel the lead reactance. These condensers C_3 and C_4 are shown in Fig. 225, the chokes L_3 and L_4 being to provide the necessary D.C. current path and prevent grid blocking.

Neutralisation of Tetrodes and Pentodes

Fig. 226 shows a normal pentode power-amplifier stage, the shunt resistance across *LC* representing the load of the following stage. The remarks following may be considered as applying to a tetrode stage. The suppressor grid is shown at zero potential,

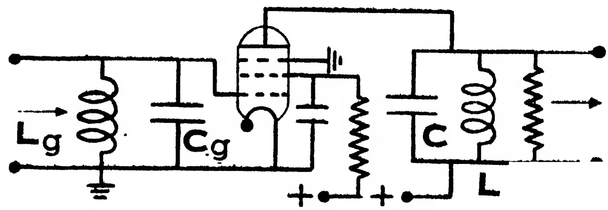


FIG. 226. Pentode R.F. Power Amplifier Stage.

under which condition most modern power-pentodes will give maximum output, and the screen is shown supplied with positive

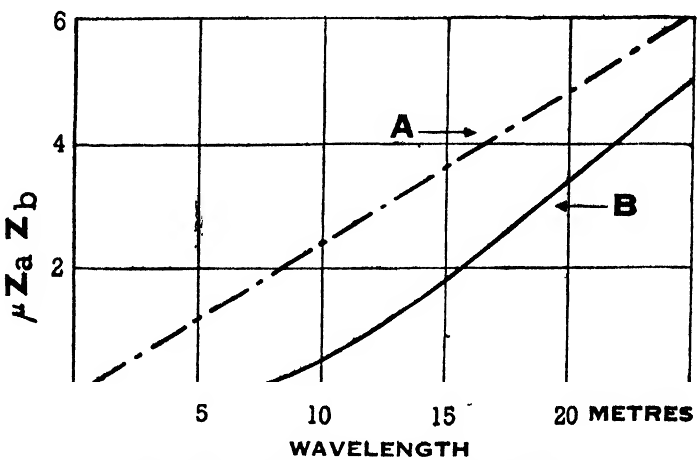


FIG. 227. Instability in Pentode Amplifier.

potential through a series resistor from the main H.T. supply, this being a normal arrangement.

It will be observed that no neutralisation circuit is shown and none should be necessary provided the mechanical construction of the system is correctly carried out, so as to remove coupling between grid and anode circuits. This involves the

screening of the resonant circuits one from the other and the correct mounting of the pentode valve within a screen.

It will be observed that since there are no neutralising arrangements provided, there is no adjustment by which instability of the pentode transmitter can be prevented as there is with a triode stage. This does not mean, however, that the pentode circuit is inherently stable. In fact, many precautions need to be taken when designing pentode circuits, and checks must be made to determine that the circuit set up is free from incipient oscillation at the driving frequency and also from parasitic oscillations.

Considering the stage such as shown in Fig. 226, it is found that if the impedances in the anode and grid circuits are too high the stage becomes unstable and the lower the wavelength the greater the instability, the connection between wavelength and stability being as shown in Fig. 227. Curve "A" is a theoretical curve which assumes that the "carry-through" capacity is mostly grid/anode and that there is no inductance in the screen, and suppressor grid leads. If the value of $\mu Z_a Z_b$, at a given wavelength, is greater than that shown on the curve, then the simplified theory states that the amplifier circuit will self-oscillate. Curve "B," the shape of which will vary with the type of valve and circuit layout, is typical for a small pentode valve suitable for about 100 watts output. The divergence in the two curves is due to the fact that there is "carry-through" via the other electrodes which cannot be truly shorted to earth potential. Thus considering Fig. 228, which shows the effective circuit, it will be seen that although the screen and suppressor grids are ostensibly earthed directly through their shunt condensers (not shown) the effective inductance of their leads will have the effect of producing a carry-through voltage between the electrode and earth and, in consequence, from anode back to the control-grid.

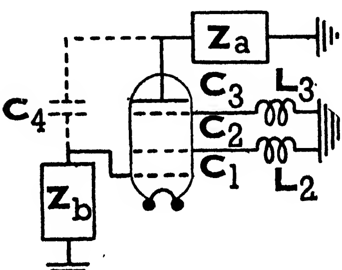


FIG. 228. Illustrating Stray Couplings in Pentode Amplifier.

For the particular valve in question, the relative values of

$\mu Z_a Z_b$ are given on the curve, and the values of $L_2 L_3$ and the inter-electrode capacitances are as under:

$$\begin{aligned} L_2L_3 &= 4 \times 10^{-15}\, H. \\ C_1 &= 25\ \mu\mu\text{F}. \\ C_2 &= 30\ \ ,, \\ C_3 &= 50\ \ ,, \\ C_4 &= 0{\cdot}05\ ,, \end{aligned}$$

Parasitic Oscillations

In addition to self-oscillation on the fundamental frequency, there are many possible "degrees of freedom" in an amplifier system, the more so the greater the complexity of the network or inductances and capacities. Such unwanted oscillations at frequencies other than the fundamental or one of its harmonics are known as parasitic oscillations, and they are more liable to occur in high-power amplifiers and where valves are paralleled.

Parasitic oscillations show up in a variety of ways: by high anode feed currents, grid currents of unusual value, circuit instability and violent voltage transients; and sometimes by an audible note when listening to the radiated wave in a tuned receiver. The ordinary anti-reaction arrangements do not stop these spurious oscillations, and as the effects are altered considerably by slight differences of circuit layout the exact arrangement to prevent a particular parasite is not calculable beforehand, but must be developed to meet each case.

It is difficult to classify parasitic oscillations, but there are one or two fairly well-defined types which may be mentioned.

(1) Parasites of a Low Frequency. The use of grid and anode chokes may lead to low frequency oscillation, due to their inductances and associated capacities. Such are prevented by suitably proportioning the values of grid and anode chokes.

(2) Parasites of a Very High Frequency. The inductance of connecting leads between valves and circuits, together with the valves capacities, may form a circuit of very short natural wavelength. Such a circuit may be found connected, not only with the high frequency valves of a transmitter, but also with the modulators, and in the event of self-oscillation considerable power may be developed in the parasite.

When the parasite occurs in longer-wave, high-frequency

circuits or in modulating circuits, it can be prevented by providing a path of high resistance in the grid-anode valve leads, and a low impedance path from grid to earth at the parasitic frequency. This is accomplished by the addition of a resistance, usually of value between 50 and 500 ohms, in the grid or anode circuit, and a condenser shunt from grid to filament.

On short wavelengths such arrangements are not possible, as they will also prevent efficient operation on the desired wave, and hence this particular form of parasite must be avoided by careful design and layout of the circuit, such that leads from valves to circuits are as short as possible or are compensated for.

(3) Parasites near the Fundamental. This parasite may cause audible modulation and in some cases is due to a choke, or choke condenser combination, or on short waves even a lead with its self-capacity to frame, having a natural frequency too near the fundamental and producing self-oscillation at a particular frequency, which beats with the fundamental.

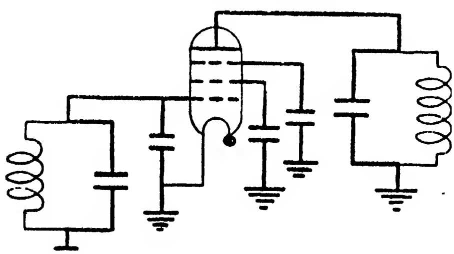


FIG. 229. Pentode Amplifier Capacitances.

If an amplifier circuit is not properly balanced, or direct coupling exists between stages, somewhat similar effects may be observed.

The presence of audible modulation, or howl, may also be brought about by a high resistance in the grid circuit of a stage which is self-oscillating. This effect, which was called originally a " squegger," is well known and need not be discussed.

In the case of tetrode and pentode valve circuits, although the absence of neutralising simplifies the circuit there are still a number of possible paths around which such oscillations can occur unless precautions are taken in the layout of components and all leads kept as short as possible. The various modes possible which are all of the ultra-short wave type can be realised by studying Fig. 229, where the heavy lines indicate

possible circuits, it being assumed that the leads shown all represent small inductances.

As the wavelength is reduced, the tendency to self-oscillate at an ultra-short wavelength increases, and influences the general shape of the wavelength-output characteristic somewhat as shown in Fig. 230. The rise of output at " B " is due

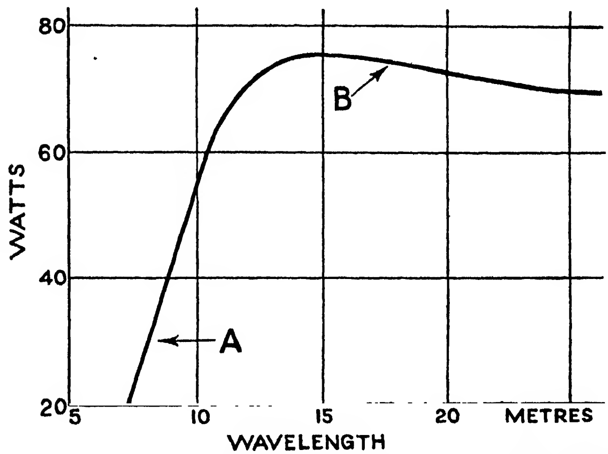


FIG. 230. Pentode Amplifier--Variations of Output with Operating Wavelength.

to a phase of reaction which assists the drive, whereas the sharp fall at " A " is due to a negative feed-back condition arising.

Comparison between Triode, Pentode and Screen-Grid Efficiency

In receiving work the triode cannot compare at all favourably with the screen grid and pentode valve as an H.F. voltage amplifier. But as has been shown earlier in the chapter, the power amplification problem is quite different and it has been seen that the conversion efficiency is dependent upon the steepness of the E_aI_a limiting edge, and in no way upon the internal A.C. resistance of the valve. To make a fair comparison, however, of the overall efficiency it is necessary to take into consideration the input and cathode circuits as well. Leaving the cathode for the moment, in the case of the triode, we have only a control-grid loss to consider, whereas in the

pentode we have the screen circuit and suppressor-grid loss to be considered as well. In the screen-grid and pentode valves, the loss in the control-grid circuit is much less than with a corresponding triode because the excursion into the grid-current region is much less. We must also remember that much of the control-grid loss is an H.F. loss, and it is important to reduce this as much as possible as it has to be supplied from a previous stage which is itself working at a conversion efficiency of perhaps 70% or less.

Regarding the screen-grid loss, which is, of course, absent in the triode valve, in a well-designed pentode the screen voltage is as low as 20% of the anode voltage and the screen loss is twofold. First, there is the actual loss in the valve itself, which is determined directly from the screen voltage and screen current, and in addition there is a possible loss depending upon how the screen is fed. If, for convenience, we feed the screen grid from the anode supply through a dropping resistance as shown in Fig. 231, we shall have an I^2R loss in this resistance. If the screen is supplied from a separate source we avoid this loss, but this method is not too satisfactory, because when the valve is modulated, the screen current will vary, and this produces distortion.

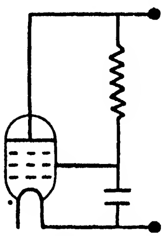


FIG. 231. H.T. Supply to the Screen-grid of a Pentode.

As regards the cathodes, in order to obtain full emission for anode and screen the cathode has to be rather larger for the tetrode and pentode valve than for a triode of corresponding size.

Considering the problem in general terms, therefore, it will

TABLE XVI

Valve.	Angle of Current Flow.	Drive Power.	Screen Loss.	Power Input.	Total H.F. Power.	Conversion Effy.	kW to Load.	Overall Effy.	Stage Power Gain.	Volts.
Triode	156°	·035	—	1·43	1·03	72%	1·0	69%	27	4,000
Tetrode	150°	·025	0·27	·97	·67	70%	·6	66%	27	4,000
Pentode	150°	·020	·114	1·90	1·43	75%	1·27	69%	63	4,000

be found that the overall efficiency of tetrode and pentode transmitting valves is not materially different from the triode counterpart as the smaller grid driving power is offset by screen and cathode losses, and the table on page 388 shows comparative figures for a triode, pentode, and screen-grid valve of similar size as regards power output and voltage of supply.

Linear Amplifiers

It was mentioned earlier in the chapter that the general shape of the input volts-output current curve is usually not linear, particularly with Class C amplifiers employing a " leak " bias, but that for telegraphy the matter is of no importance.

In modulated-carrier transmissions, as telephony, and particularly with single-sideband work, where the inter-modulation frequencies must be kept down to avoid cross-modulation, the linearity of any grid-modulated stage and all subsequent stages is most important. Assuming static bias is used on the stages in question, non-linearity of the input-output curve is still evident, due partly to the curvature of the valve characteristics, and is worse the smaller the angle of current flow. It is mostly due to the fact that since grid current flows only for a portion of each H.F. cycle and the fraction varies under modulation conditions, a varying load is thrown upon the driving valve, which tends to affect its regulation.

Although we can design a Class C linear amplifier, it is simpler and more usual to employ Class B, and where two Class B valves are used in push-pull, since current flows for 360° of the cycle, we get a Class A stage. The static bias will, of course, be adjusted to a point just above cut-off, i.e. so as to give an approximately straight, overall characteristic. A number of ways have been tried to overcome the effect of grid-current unequal loading. Most circuits in the past have been designed with a low value resistance in shunt with the grid-cathode of the amplifier valve. This means, of course, increased output from the driving valve since the resistance current (and power) is many times that required for the grid. Such a method is therefore uneconomical and may require an input power of some ten times or more that necessary to drive a similar amplifier under telegraph conditions.

Another method is to run the driving valve as a cathode follower (see page 521) with the amplifier grid circuit as the cathode load, but this necessitates a greater voltage from the driving valve and, since the driving valve is above earth and at H.F., complicates the design.

An alternative circuit to either of the above and one which avoids the disadvantages, employs a high-impedance driving valve of the screen-grid type operated on the flat slope, working through an impedance-transforming network (simulating the conditions in a quarter wave line) into the grid of the amplifier, one such circuit being as shown in Fig. 232. Here, a pair of screen-grid valves, with tuned anode circuit, are driving a pair of neutralised triodes with grid and anode circuits both tuned

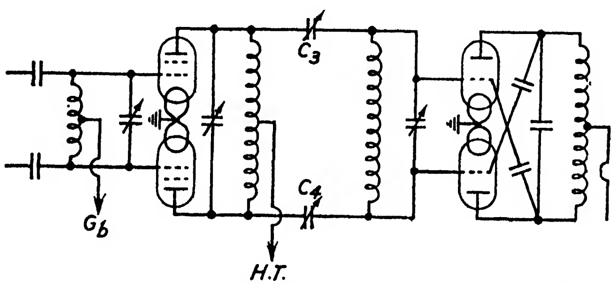


FIG. 232. Linear Amplifier with Impedance-Transforming Network.

(the grid-bias point to the centre of the triode grid coil has been omitted), the condensers C_3 and C_4 each being adjusted to equal a value of $\frac{-X}{2}$, where X is the characteristic impedance of the equivalent quarter wave network. How such a circuit works can best be seen by considering the circuit discussed on page 289, namely the grid loaded up by a low resistance shunt. We may regard such a circuit, looking back from the grid-cathode terminals, as if the grid was driven from an effective generator of impedance low compared with the grid-cathode circuit. The greater the difference between the high variable impedance of the latter and the constant low resistance of the effective generator, the better the regulation of the driving circuit. A similar condition will be obtained if an actual high impedance generator is employed operating through an appropriate quarter-wave network, since the low resistance may

thereby be removed and the power loss saved. Using such a circuit, as shown in Fig. 232, linearity of input to output is good and the power relationship of input to output is 1/50, i.e. a gain of 7 db, as good as with telegraph conditions without the swamping resistance across the grid.

Inverted Amplifiers

Before leaving the question of an amplifying stage, we propose to describe briefly an original form of amplifier devised by Standard Telephones and Cables Ltd., and used at the B.B.C. Station at Daventry. This is " series connected," or " inverted " as it is called. A simplified diagram of connections is shown in Fig. 233 and an equivalent circuit in Fig. 234.

The unusual feature is seen to be that, on the input side, the grids are earthed and the filaments are at the high-frequency

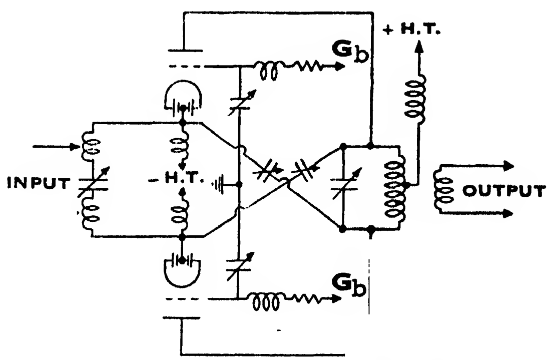


FIG. 233. Circuit of Inverted Amplifier.

potential. In the actual amplifier the grids are earthed through series condensers which neutralise the inductive reactance of the grid leads and these condensers must therefore be adjusted for each change in operating frequency.

It will be seen that if the grid leads are effectively earthed to high frequency then they form a screen between input and output circuits similar to the screen in a tetrode or pentode valve. The tendency to self-oscillation is thereby reduced and when neutralising condensers are necessary they will be very much smaller than in the conventional triode circuit, which assists in the design of high-power, short-wave amplifiers. The capacity across the anode tank circuit is also less with the

inverted arrangement because of the small size of balancing condenser.

In Fig. 234 one-half of the push-pull arrangement is drawn in its simplest form where A, G, C are the anode, grid and cathode of the amplifier valve, and in Fig. 235 is shown the corresponding curves for one cycle of potential between them. E_{CE} is the voltage between cathode and earth from the exciter stage, E_{GC} the voltage between grid and cathode of the amplifier, and E_{AC} the amplifier anode voltage, from which it is seen that the amplifier and exciter voltages are in phase, but as usual the amplifier grid and anode voltages are in anti-phase. Further, that the amplifier valve and exciter circuit are in series across the output circuit and the phase relationships are such that the drive supplies some power to the output stage. The exciter circuit has, therefore, to be of larger

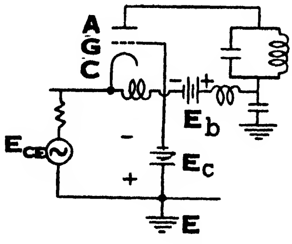


FIG. 234. Equivalent Circuit of Inverted Amplifier.

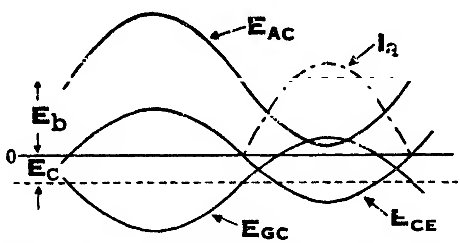


FIG. 235. Phase Relationships in Inverted Amplifier.

rating than in the conventional design but some of the extra power is utilised in the output.

The alternating component of the amplifier anode current evidently flows through the impedance of the exciter circuit and this provides negative feed-back of an amount which depends upon the adjustment of the exciter circuit.

When such an arrangement is modulated it is necessary to modulate the exciter stage as well as the amplifier because of the fact that some of the drive power appears in the final output.

Multi-stage Amplifier

Thus far we have discussed only a single stage of amplification, but except for a very small transmitter, more than one stage will be employed, the number used depending upon many factors. In general more stages will be required with triodes than pentodes or screen-grid valves, since with the latter the power gain per stage is greater. The type of drive used will also influence the matter since if a drive can be allowed a wide tolerance of frequency-drift it can be designed

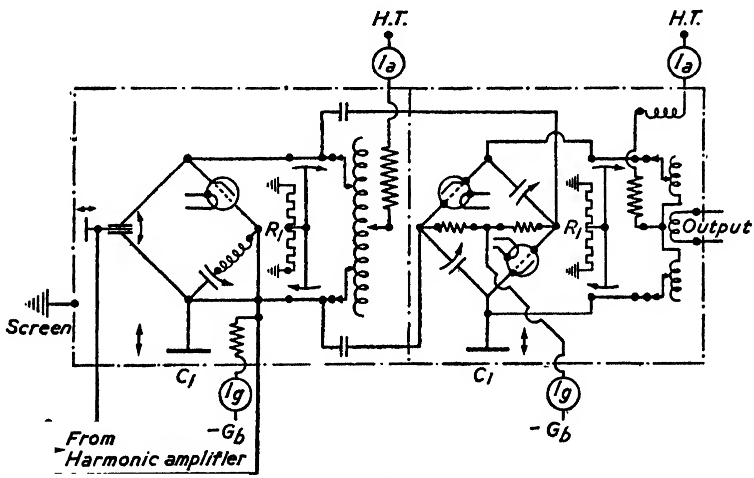


FIG. 236. Two Stage R.F. Amplifier.

for a fairly large power valve, but with a so-called constant-frequency drive the output from the drive unit will be very small. The wavelength and range of wavelengths also has some bearing, and the type of modulation that is to be employed has also to be considered. As a very general rule, small sets will have one stage, and medium and large sets, two or three, but rarely more, these stages all operating on the same and final frequency radiated. A typical circuit diagram is shown in Fig. 236, which shows a two-stage main amplifier, as distinct from the pre-amplifier and frequency-multiplying circuits to be discussed shortly. The stages shown give examples of single and double valve bridge circuits and little explanation is required. The capacity plates marked C_1 to

earth give a fine control to level up the heating of the two push-pull valves, and the non-inductive resistances R_1 are to hold the centre point of the condenser plates to earth.

Circuits for Frequency Multiplication

The use of frequency multiplication is very common on short waves because it is often much easier to obtain certain effects at a lower frequency and multiply the resultant up, rather than obtain the same effects at a high frequency directly. Examples are found in constant-frequency drives, single sideband transmission and frequency modulation.

Thus for a short wave transmitter operating on a frequency range from 5 to 20 Mc/s, a drive frequency of 2 Mc/s might be

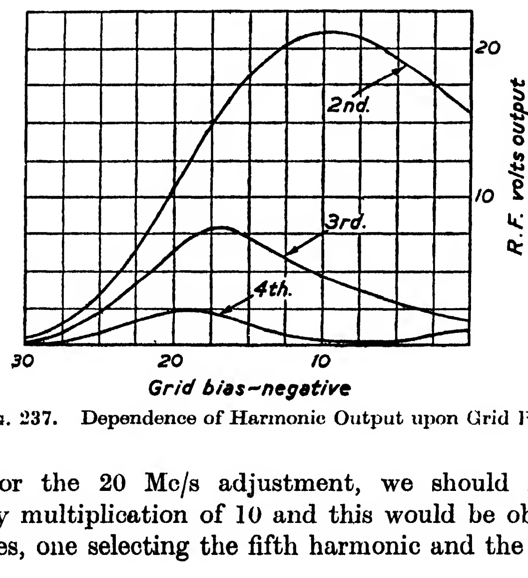


FIG. 237. Dependence of Harmonic Output upon Grid Bias.

used. For the 20 Mc/s adjustment, we should require a frequency multiplication of 10 and this would be obtained in two stages, one selecting the fifth harmonic and the other the second.

In connection with this subject one important point, which is often missed, is that to produce harmonics one does not need the original frequency source to be rich in harmonics. In fact in most cases it is desirable that the original frequency should be fairly pure in waveform, the production of harmonics being solely a function of the asymmetry of the circuits through which the wave is passed. A rectifier valve naturally lends

itself for this purpose, and this is the simplest form of frequency multiplying device.

Thus if a valve be biassed to the cut-off point and a frequency of f_0 applied to the input grid, current can only flow for each alternate half-cycle, and the waveform of the feed-current will be a series of half-sine waves. Such a waveform is rich in harmonics, and hence the tuning of an output circuit of high Q to these various harmonic frequencies will lead to the output being energised at the harmonic frequency and if it is selective enough each harmonic can be individually selected. Actually a bias to cut-off is not very effective, even for the energising of a circuit of $2f_0$, a bias to rather more than cut-off being more efficient, and the higher the harmonic required the greater the negative bias necessary, Fig. 237 showing the connection between bias and harmonic output at various multiple frequencies, from a frequency multiplier employing a single triode.

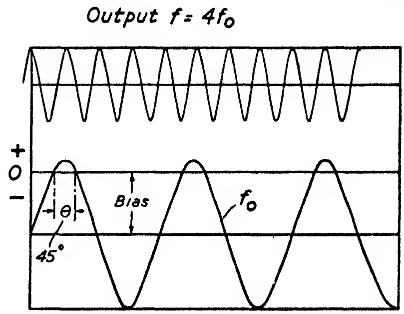


FIG. 238. Generation of a Harmonic from a Sinusoidal Fundamental Frequency.

We can think of the problem in terms of angle of current flow. Because of the nature of the circuit, the harmonic output circuit can only be energised over one occasional half-cycle of its multiple frequency, the bias being such that with a fundamental input voltage of frequency f_0, the angle of current flow θ is reduced to a value equal to a half-cycle of the multiple frequency for the most efficient result. Thus if a fourth harmonic is required, since the output circuit is only impulsed once every fourth half-cycle, the bias must be sufficiently negative for θ to be reduced to one-eighth of 260°, i.e. 45°, as shown in Fig. 238.

Quite clearly, the higher the harmonic the smaller the useful part of the impulsing curve, and the greater must be its input to obtain any given output. Actually, of course, the input is more or less constant in amplitude, in consequence of which the higher the harmonic selected the smaller the output usually

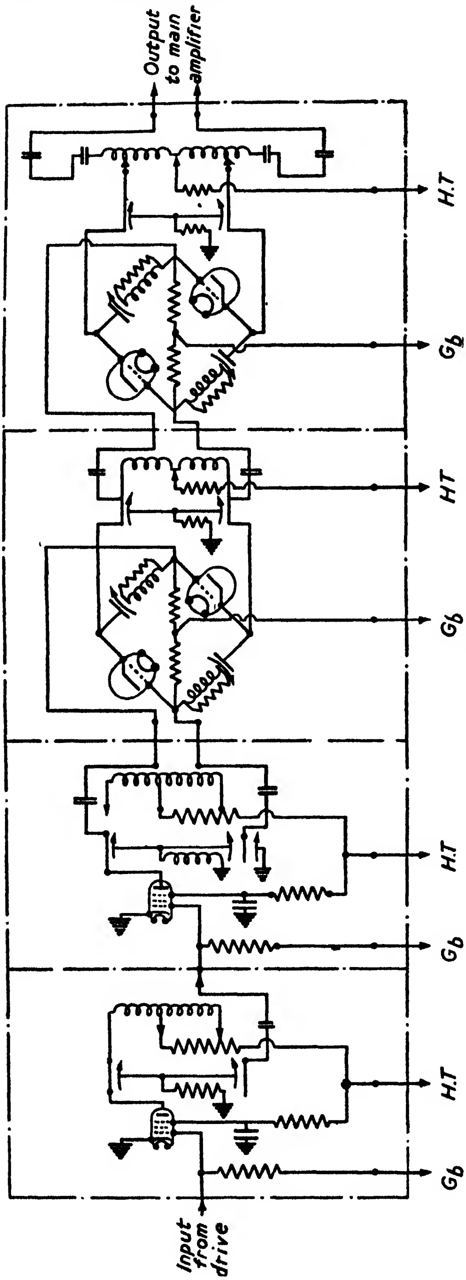


FIG. 239. Harmonic Amplifier.

obtained. As has been mentioned previously, the push-pull circuit cannot give any even harmonic output, although if it is desired to use push-pull circuits to preserve symmetry, frequency doubling can be achieved by disconnecting one lead of one valve filament. The disconnected valve then acts simply as a capacity to preserve the bridge balance, the active valve being used for doubling.

An efficient doubling circuit can also be obtained using a two-valve circuit with differential input, and the anodes paralleled to the output, both valves of course being backed off to more than cut-off. In this case the output circuit obtains an impulse every cycle of its double frequency, instead of every other cycle, as with a single valve circuit.

Triodes, tetrodes and pentodes are all employed in harmonic amplifiers, triodes more usually in the higher-power stages. Where triodes are used neutralisation will be necessary. It should be remembered that in a multi-stage harmonic amplifier, to obtain the requisite number of harmonics of high value, the changes will be rung on a few low order harmonics in each individual stage, and any one stage will be operating at times on the fundamental of its input frequency. Hence with a triode, neutralisation would be essential for such a case, but it is found that, apart from this, triode circuits still have a strong tendency to oscillate even when the circuit is tuned to a harmonic frequency, due to feed-back of harmonic energy.

Fig. 239 shows a typical harmonic amplifier circuit employing both pentode and triode valves. Note the de-coupling arrangements and the neutralisation of the triode stage, which would normally operate on the final frequency. Thus if a frequency multiplication of 12 times was required, the first valve would probably be designed to select the fourth harmonic, and the second valve the third harmonic of its input, the resultant twelfth being passed to the input of the triode stage.

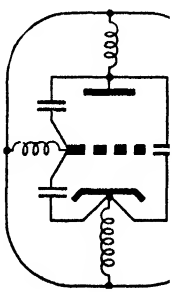
Valves for Ultra-High Frequencies

It is clear that the successful development of short and ultra-short wave apparatus depends largely upon the valve design, which has had to go hand-in-hand with the circuit technique.

As we endeavour to use valves at increasingly high frequencies, effects which are negligible at lower frequencies may become very prominent. We have seen already that the capacitances between the various electrodes make their presence felt, even at low radio frequencies. The leads between the seals and actual electrodes must have inductance and this, whilst normally too small to matter at the lower radio frequencies, becomes important when the valve is to work at many megacycles. Even the triode, therefore, becomes a complicated network such as that shown in Fig. 240a, from which it is clear that the p.d.'s between the electrodes themselves at V.H.F. may be very different to those applied between the external connections. At a sufficiently high frequency it may be impossible to get phase relationships correct for efficient amplification.

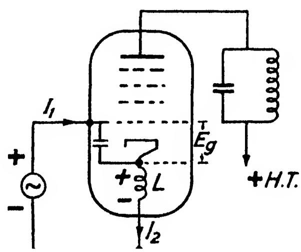
It will be seen that if we connect two electrodes by a direct lead a resonant circuit is formed by it and the capacitance and inductance inside the valve and the frequency to which this tunes represents a limit beyond which we cannot use conventional circuits. In the discussion on oscillators for U.H.F. in the following chapter it is shown that lines can be used to advantage, instead of the conventional *LC* circuits and after reading this chapter the application to amplifiers will be clear.

Because much of the circuit capacitance is within the valve, it follows that large radio-frequency currents will pass through the valve and the seals will have to be massive, not only to



(a)

FIG. 240 (*a*). Triode, showing Lead Inductances and Electrode Capacitances.



(b)

FIG. 240 (*b*). Illustrating Effect of Cathode-Lead Inductance in a Pentode.

carry these, but to reduce the temperature at the glass and, of course, to keep the lead inductance to a minimum.

When valves are used at U.H.F., it is found that the grid/cathode resistance falls greatly. This is of especial importance in voltage amplifiers for receivers, but also means that the driving circuit of a power amplifier will require to have a greater output, until a limit is reached where the power of the drive circuit is as great as the output from the amplifier. This reduction of input resistance is due to two causes, which will now be discussed.

Consider first the effect of the cathode inductance. If we are not to get very cumbersome equations it will be necessary to simplify the problem somewhat, but the conclusions reached will apply to practical cases. Let us consider a pentode

circuit, as in Fig. 240b, in which the grid is supposed to be negative throughout the cycle, the alternating screen grid-current is neglected, the anode circuit is in tune and phase shifts due to L and to transit time are also neglected.

Then
$$I_2 = g_m E_g \quad (I_2 \gg I_1)$$
$$E_g = -\frac{jI_1}{\omega C} \, (C = \text{grid/cathode capacitance})$$

Hence
$$E = -\frac{jI_1}{\omega C} + j\omega L I_2 = -\frac{jI_1}{\omega C} + j\omega L g_m \left(-\frac{jI_1}{\omega C}\right)$$
$$E = -\frac{jI_1}{\omega C}(1 + j\omega L g_m)$$

or
$$I_1 = \frac{j\omega C E}{1 + j\omega C L g_m}$$

Input admittance
$$= \frac{I_1}{E} = \frac{j\omega C}{1 + j\omega L g_m}$$
$$= \frac{j\omega C(1 - j\omega L g_m)}{1 + \omega^2 L^2 g_m^{\ 2}}$$

If $\omega^2 L^2 g_m^{\ 2} \ll 1$

Input admittance $= \omega^2 L C g_m + j\omega C$, so that the cathode lead inductance has introduced a conductance of $\omega^2 L C g_m$.

There are practical difficulties in reducing the cathode inductance since a length of lead is necessary to ensure that the seal is at a much lower temperature than the cathode itself. In U.H.F. circuits, therefore, it may be better in some cases to use circuits in which the cathode is not common between input and output, such as the inverted amplifier already discussed.

Electron Transit Time

A second factor which reduces the grid/cathode resistance is the transit time of the electrons crossing from cathode to anode. A quantitative analysis of this problem, applicable to practical electrode systems, is difficult but the following explanation shows the way in which the reduction of resistance comes about and also the factors upon which it depends.

Let us first consider what would happen in a diode if a single electron crossed over from cathode to anode. All the time it

was crossing, work would be done on it, and this requires that there should be an anode current so that the electrical energy put in (the product of anode voltage, anode current and time) may equal the work done. In a given field this will be proportional to the distance travelled in the time, that is, to the velocity of the electron. When the electron arrived at the anode, the current would cease and the kinetic energy of the electron would be converted into heat.

Of course, in any actual valve, myriads of electrons are crossing at the same time and the current is nearly constant, though it does fluctuate, giving rise to the shot noise (discussed on page 555). The mean anode current can be calculated from the charge brought to the anode in a second and for this purpose it is not necessary to realise that the current flows all the time that the electrons are in flight between cathode and anode. When the time of flight of the electrons becomes comparable with the period of the alternating voltages applied to the electrodes, however, it does become necessary to take this into account.

It will be seen, also, that currents flow into or out of electrodes whenever electrons are approaching or receding from them, whether they ever reach the electrodes or not. This fact is frequently of importance in special valves used for ultra-high frequencies.

Consider now a triode valve whose grid is sufficiently biassed negative by D.C. to prevent D.C. grid current at any time, and assume for simplicity the valve has no anode load, but that a D.C. anode feed is flowing. From our previous discussion we realise that the electrons approaching the grid will tend to cause a current to flow in it, whilst the electrons leaving towards the anode will tend to cause a current flow in the grid in the opposite direction. Under steady conditions, these currents will be equal and balance. It is actually unlikely that there will be exactly equal numbers of electrons on either side of the grid in an ordinary valve, but the product of electron density and velocity will be constant, and this will result in equal and opposite currents in the grid making the net current zero.

Let us now apply a small, low-frequency voltage between grid and cathode. When the grid is made less negative by this

voltage, then the number of electrons passing from cathode to anode will increase, but we can assume this increase to take place over the whole cathode-anode space at the same instant in the cycle since the frequency is low. Both components of grid current will therefore increase and will balance, as before, and this will also be true for all other instants in the cycle, so that the grid current is still zero. This condition is shown by the two current curves (shown dotted) i_1 and i_2, Fig. 241, which are in phase with e_g, one being reversed in sense, since current directions are *to* and *from* grid.

If the frequency is increased till a cycle is comparable with transit time, however, conditions change. At an instant when the grid is being made less negative, the number of electrons

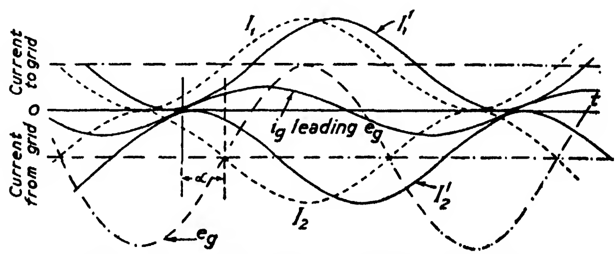


FIG. 241. Illustrating Effect of Electron Transit Time.

leaving the cathode increases immediately but this condition will travel through the valve, taking a time which is not negligible compared with the period of the alternating voltage. Hence the component i'_1 of grid current due to electrons in the cathode-grid space and i'_2 due to those in the grid-anode space, may be as shown in Fig. 241 (full lines), and will have a resultant, i_g.

This current is seen to have a large phase-angle α with respect to the grid voltage but not to be entirely in quadrature with it, and the greater the transit time the more this current will depart from quadrature. We can see from energy considerations that the current must have an in-phase component. When the grid voltage is increasing more work will be done on electrons approaching the grid (there being more of them) than will be given up by electrons leaving, so that there is a net supply of energy from the grid circuit. Similarly, during other portions of the cycle.

The quadrature component of I_g results in an increase in the effective capacitance between grid and cathode, compared with the value at low frequencies, but this is of little practical importance. The in-phase component, however, reduces the input resistance and this can produce heavy damping on the input tuned circuit, in addition to that which may be caused by cathode inductance.

Although our discussion is qualitative we can suggest from it the way in which we think the various factors will affect the value of the input resistance R_i. If T is the transit time, f the frequency and g_m the mutual conductance, then we should expect I_g to be proportional to T and f, since I_g would be negligible if either of these were very small. Also, I_g would be proportional to g_m since this determines the alternating current for a given E_g.

Now the in-phase component of I_g is given by $I_g \sin \alpha$, or, approximately, by $I_g\alpha$ if α is small. But α is also proportional to T and f and hence the in-phase component of I_g is proportional to $g_m f^2 T^2$ or

$$R_i = \frac{k}{g_m f^2 T^2},$$

where k is a constant depending upon the geometry of the valve and electrode potentials. This relationship is confirmed by the more complete analyses of North, Llewellyn and Benham.

It will be seen that for a given valve operated at fixed potential, $R_i = \frac{k}{f^2}$. This varies in the same way with f as the effect due to cathode inductance, and they are not readily separated, therefore, in measurements of input resistance.

Another result of transit time, which may be of importance in the operation of a power amplifier at U.H.F., is that, although the maximum number of electrons leave the cathode when the grid is at its maximum positive value, yet they do not arrive at the anode when the volts there are a minimum (examine Fig. 216, where transit time is assumed to be zero). As a result the losses are increased, although the effect can be partly compensated for by detuning the anode circuit slightly.

In a certain water-cooled valve designed for short-wave working, the transit time from cathode to grid is $0{\cdot}65 \times 10^{-9}$ secs., that from grid to anode $1{\cdot}3 \times 10^{-9}$ secs., or the total

time from cathode to anode 2×10^{-9} secs. At a frequency of 50 Mc/s this represents one-tenth of a cycle.

Valve Types

The triode has for many years been, and still is, a most efficient type of valve for use in radio transmitters. Special designs of triode valves have been evolved (see page 408) to meet the particular requirements of high and ultra-high frequencies; and conversion efficiencies can be obtained with a modern S.W. triode as high, or even higher, than normally obtained at lower frequencies.

Beam-tetrode and pentode valves for transmitting work are, however, also in common use in the medium and lower power sets. The beam-tetrode is a screen-grid valve designed so that secondary emission from the anode is eliminated by the concentration of the electrons into a beam, which increases the space-charge effects existing in the anode/screen space and ensures that a potential minimum is created near the anode surface even at low anode-voltages. This effect is achieved by the aligning of the grids in the valve and the inclusion of two beam-forming electrodes, parallel to the screen supports and placed between the screen and anode. Such valves are not usually more efficient than the triode, but where flexibility is required they have certain advantages because their low inter-electrode capacity eliminates the use of neutralisation arrangements and this simplifies the circuit layout. In the case of the pentode valve, it is possible to carry out effective low-level, 100% modulation by using the suppressor-grid as a control electrode.

The design of beam-tetrode and pentode valves for transmitting work calls for features which are somewhat different from those met with on similar valves developed for receiving work because the avoidance of power loss is an important factor, and the seals of the various electrodes will be required to carry considerable high-frequency currents, and must be designed to have the minimum possible resistance and inductance. In a modern transmitting pentode a conventional seal will be used for the control electrode g_1, but the connecting wires of the screen-grid g_2 and the suppressor-grid g_3 are of multiple form and are taken away through

the base of the valve. By this means, the inductance of the leads is reduced to the minimum possible, the current distribution is made uniform and the current-carrying capacity large.

One of the advantages of the pentode valve is the ability to use suppressor-grid modulation (see page 532) and this necessitates certain guiding rules in designing the valve. If its amplification factor μg_3 is made high, then only a few volts swing are necessary to control the anode current fully, but the peak anode-current can only be reached with positive volts on g_3, the suppressor grid. This is an undesirable feature for modulation purposes, as grid current will be taken, which would load and distort the modulation output. If the amplification-factor is made too low, the peak anode-current can now be obtained with zero or even negative voltage on g_3, but the output is lower than with the previous design and a considerable voltage swing on g_3 is required to fully modulate the anode current. A further consideration is that the design of g_3 also affects the screen (or g_2) current, the higher the amplification of g_3 the less the screen current, and because only small changes of g_3 will fully modulate the valve these changes will be accompanied by only a small change of screen current. Thus when the valve is designed for suppressor grid modulation a compromise is adopted, the peak anode current being obtained with g_3 at zero voltage or very slightly positive.

FIG. 242. Air-cooled—S.W. Valve.

Another important feature in the design of both the beam-tetrode and pentode valves is the reduction of the screen losses to the minimum possible. As has been mentioned, this

can be helped by keeping up the amplification-factor μ_{g3}, but mostly it is brought about by the principle of alignment, i.e. the grid structures are not wound random fashion, but correctly aligned one with the other so that each g_2 wire is exactly behind a g_1 wire, and similarly with each g_3 wire. A typical screen grid and pentode transmitting valve is generally designed so that with a screen-potential of some 20% of the anode-voltage it is possible to obtain peak anode emission when the anode potential is reduced to some 15% of the working average potential, i.e. giving similar conditions to those obtained with triode valve.

The effect of inductance in the screen-grid circuit has already been discussed (page 384). It is evident that it will not be possible to tie down the screen grid completely to earth potential at U.H.F. because of the inherent inductance in its internal lead. This will cause a negative resistance to be put back into the input circuit, thereby raising the input resistance and partly counteracting the effects of cathode inductance and transit time.

From a consideration of the above effects it will be seen that all valves to operate on U.H.F., whether for voltage amplification in receivers, power amplifiers in transmitters, or as oscillators, should be designed with short, well-spaced connections to the electrodes. Also, the electrodes should be placed as close together as possible in order to reduce transit time. This, however, will result in a large increase in inter-electrode capacitances unless the area of the electrodes is also reduced.

It can be shown that if two valves are constructed, having electrode systems of the same shape but in the one case the linear dimensions are $1/n$th that in the other, then the two valves will have similar values of r_a, μ and g_m. Inter-electrode capacitances and electron transit time between the various electrodes will, however, be divided by n. The anode dissipation and cathode emission of the smaller valve will only be $1/n^2$ of those of the larger valve because these quantities are constant per unit area.

It is clearly very desirable that valves which are alleged to be the same should, in fact, have the same characteristics. Even in the manufacture of the larger valves the tolerances in dimensions have to be made very small and a very careful check

kept on the manufacturing processes if a reasonable degree of uniformity is to be obtained. These difficulties are, naturally, greatly increased if the valve is reduced in size, the tolerances also being reduced to $1/n$ their former value.

Transit time can be reduced by employing higher anode voltages so that the velocity of the electrons is increased. For a given electrode system this results in a greater anode current, and, if saturation is not to occur, the emission of the cathode must be increased. In fact, it can be shown that the required cathode emission increases as the cube of the frequency.

If the cathode-grid spacing is made very small, then the grid will get very hot and may give trouble from secondary emission and may also suffer damage.

A consideration of the above factors will reveal that the difficulties will greatly increase as we attempt to construct valves for higher powers. If the electrodes are to remain small, then the cathode emission per cm will have to be very high, and the anode dissipation per cm will also be high, in addition to the difficulty of grid heating and of heavy seal currents.

Construction of Valves

The general requirements for valves, particularly when they are to be used at very high frequencies, having been discussed, it remains to consider briefly the actual construction employed. One valve of each main class will be used as an illustration, it being understood that there are numerous types by several makers.

The air-cooled glass valve was the first type and is still the cheapest type to produce and suitable for low-power transmitters. Clearly, all the power dissipated at the anode has to be got rid of by radiation and the wire seals have to carry the H.F. currents, which may be heavy in the case of the higher frequencies.

The anodes are usually of molybdenum or carbon and in many cases the glass envelopes have all metallic deposits washed out after manufacture. At higher radio frequencies, eddy currents circulate in these deposits and produce hot spots. If the washing-out process has not been carried out, it is desirable to enclose the valve with a copper strip or gauze

jacket. This has the effect of equalising the potential distribution over the glass surface, it not being necessary for such a jacket to be connected to earth.

For valves up to some 200 watts dissipation it is, generally speaking, not necessary to use forced air circulation or to raise the filament voltage in steps when switching on. Thoriated tungsten filaments are often used.

Radiation valves are also used in which the envelope is of silica, thus enabling higher temperatures to be reached, with correspondingly higher anode dissipations, which may reach several kilowatts. It is usually necessary to cool the seals of such valves with an air blast.

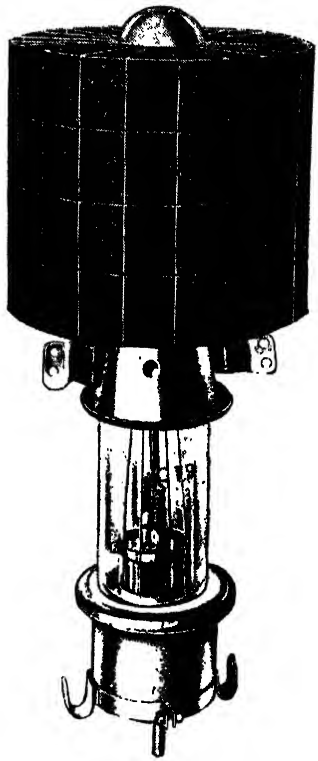
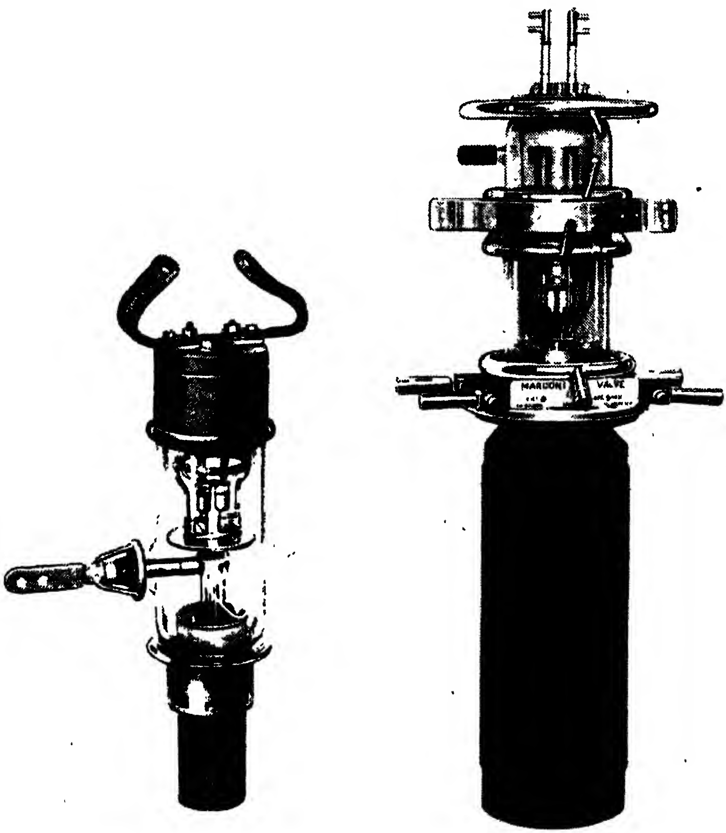


FIG. 243. Forced Draught Air Cooled Valve.

At a later stage seals were developed which enabled valve envelopes to be built up of metal and glass. This made it possible for the anode to be part of the envelope, so that it could be cooled externally. Convection air cooling is suitable for smaller sizes (up to about 500 watts dissipation). Louvre attachments are frequently fitted to the anode in order to encourage the circulation of air past it. If forced-air cooling is used the permissible dissipation may be increased by about 50%.

The use of this construction has great advantages for very high-frequency operation, beside the increased dissipation. The anode can be connected directly to its circuit, instead of through a rather long, thin lead, the inductance of which will cause trouble and the current-carrying capacity of which will be small. In some valves the grid is also brought out through an annular seal instead of merely through a wire seal.

Valves of the highest power use water cooling for the anode and valves capable of dissipating 150 kW are used in some short wave transmitters. An interesting feature of such valves is



C.A.T.15. C.A.T.17.

FIG. 244. Water Cooled Valve.

that the water-column thickness between anode and jacket is only about $\frac{1}{4}$ inch.

Tungsten filaments are invariably used in the large, water-cooled valves, as thoriated or oxide-coated cathodes will not stand up to the operating conditions.

The large transmitting valves are very expensive and their useful life is that of their cathode. It has always seemed an

attractive proposition to build a valve of steel and porcelain which could be readily dismantled and a new cathode fitted. Such a valve would not maintain a sufficiently high vacuum and pumps would have to be provided with the valve. In recent years, however, such pumps have become very reliable and are fitted to mercury-arc rectifiers in unattended substations.

Demountable valves have been used to a very limited extent, however. One such valve, giving an output of 500 kW, has been used on the Rugby long-wave transmitter and early metre-wave radar transmitters employed such valves.

Pentode valves of all the above classes are also available.

Notes on General Design

The design of short and ultra-short wave transmitting apparatus calls for special attention both to the materials used and to the design and layout of the components. All circuits must be carefully screened from one another by good-quality copper, brass or aluminium sheets, carefully bonded to each other and earthed to a definite, earth bus-bar. If this is not done, not only will there be interaction between the circuits but the proximity of any R.F. circuit to surrounding dielectrics, such as wood or plaster walls, will increase the losses very considerably, even if no dangerous heating is set up in the dielectrics themselves.

The R.F. circuits must be designed so that as few insulating materials are used as possible and where they must be incorporated they should be so designed as to be outside the influence of strong electric fields, if this can be arranged. Unless acting as a transmission line all connecting leads should be of ample dimensions and kept as short and straight as the design permits. Apart from the general screening of the H.F. apparatus, supply leads to filaments, grids, etc., should be run in screened compartments, or tubing, as near to their final connecting point as possible, and adequate de-coupling for all such circuits must be provided. As regards the tuning coils themselves, continuously variable inductances are coming into use more and more as the design of sliding contacts improves at high frequencies, but many sets still employ fixed inductances. In this case, either separate coils designed for each waveband are used,

or an even number of single loops, which can be connected in various ways. Thus they would be joined all in series for the longest wavelength group, all in parallel for the shortest, and in series-parallel for intermediate groups. Such a design automatically increases the conductor size as the wavelength decreases and current increases.

With high powers and high voltages it is necessary to avoid all sharp corners and edges to prevent corona. At very high frequencies the energy supplied to the air by corona is much greater than at lower frequencies, general ionisation of the air around takes place and a flash over occurs which is more in the nature of an arc at such frequencies. The arc may flash over to another electrode as with such discharges at lower frequencies, but it will often be found to arc out into space, when it is known as a torch discharge. To avoid these discharges, all conductors carrying H.F. current should be designed with rounded edges, and this means, with condenser plates, either making them heavy, or building up a hollow section, an example of the latter type of construction being shown on page 689.

Copper tubular coils are suitable, or flat copper strip with the flat surface parallel to the axis of the coil. A very convenient design is the square-shaped inductance used by Franklin (see Fig. 403), as this lends itself not only to mechanical design, *but to easy cross-connecting and fine adjustment of individual* turns, without giving any detrimental end-effect.

The spacing of turns is usually not greater than twice the diameter (or width) of the turns, and very little alteration of inductance is obtained by altering the spacing, because the end-to-end capacity is of as much consequence as the increase of mutual inductance between the turns.

Dielectrics for High Frequencies

A considerable number of insulating materials suitable for use at high frequencies have been produced in recent years as many of the materials which are efficient at low frequencies have far too high a dielectric loss at high frequencies, since this tends to be a constant per cycle of applied voltage.

The effects of a high dielectric loss are, of course, that the effective resistance of the circuit concerned is increased and the dielectric may become distorted if not destroyed. In the

case of materials to be used as insulating supports, the main requirements are mechanical strength, low dielectric loss and high dielectric strength, the dielectric constant being relatively unimportant. They should be non-hygroscopic. Amongst suitable materials is "Mycalex" (made from mica and glass), as this can be obtained in a variety of forms, and lends itself to drilling, milling, and ordinary machine processes. Other materials are "Frequentite" and "Calan," which are manufactured in the same way as other ceramic materials, the former being made from steatite and the latter from finely divided mica. Of the numerous synthetic resin insulators "Trolitul" has excellent properties, and like other plastics can be moulded into complicated forms. Porcelain, although it is very uneven in character, is still extensively used, particularly in places outside the H.F. field, but only those types which contain no filling material except lead are really suitable.

When dielectrics are wanted for building up condensers, a low dielectric loss is evidently desirable, mechanical strength is not so necessary, but a high dielectric constant is useful in order to reduce the size of the condenser, this being especially useful because the small size reduces the inductance of the leads and plates themselves.

In short and ultra-short wave transmitters, the tuning capacity is almost always a low-capacity, variable air-condenser, but for blocking condensers mica sheets of good quality are

TABLE XVII

Material.	Dielectric Constant.	Power Factor.	Dieletric Strength kV/mm
Calan	6·5	·0004 at 10^7	40
Frequentite	6·0	·0008 at 10^7	50
Mycalex	6·0	·003 at 10^6	14
Mica	7·0	·0002 —	50
Trolitul	2·5	·0003 at 10^6	30
Porcelain	5·5	·008 at 10^6	—
Fused quartz	3·8	·0002 —	20
Ebonite	3·0	·009 at 10^6	150
Kerafar	80·0	·001 at 10^6	—

Extracted from the "Journal of Scientific Instruments," Vol. 15, page 217. Values obtained by Dr. Hartshorn of the N.P.L.

still much used. Recent Continental research has produced some remarkable materials having enormous dielectric constants. These are being increasingly used in receiving components and will no doubt become more extensively used in transmitting equipment. The properties of some dielectrics are tabulated on page 411, and the table shows the notable advance of these materials over the rubber compounds such as ebonite, the standard insulating material of a few years ago.

Selected References

(1) PRINCE. "Vacuum Tubes as Power Oscillators." *P.I.R.E.*, vol. 11. 1923.
(2) TERMAN AND ROAKE. "Calculation of Class C Amplifier." *P.I.R.E.*, vol. 24. April, 1936.
(3) WAGENER. "Simplified Methods of Computing Performance of Transmitting Tubes." *P.I.R.E.*, vol. 25. 1937.
(4) GREEN. "Comparison of Parallel and Series Coupling Circuits for Transmitters." *Marconi Rev.*, 68 and 69. 1938.
(5) STRUTT AND VAN DER ZIEL. "Causes for Increase of Admittance, etc." *P.I.R.E.*, vol. 26. August, 1938.
(6) FERRIS. "Input Resistance of Vacuum Tubes as Ultra-High Frequency Amplifiers." *P.I.R.E.*, vol. 24. 1936.
(7) NORTH. "Analysis of the Effects of Space Charge on Grid Impedance." *P.I.R.E.*, vol. 24. 1936.
(8) PICKEN. "Development of Wireless Transmitting Valves." *J.I.E.E.*, vol. 88. 1941.
(9) GAVIN. "Valves for Use at Ultra-Short Wavelengths." *W.E.*, vol. 15. 1939.

CHAPTER XI

OSCILLATORS AND CONSTANT FREQUENCY OSCILLATORS

ALTHOUGH the discovery of the utility of short waves greatly extended the useful "wireless spectrum," at the same time it so greatly increased the usefulness of wireless communication that the new frequency bands were quickly filled up. Since the modulation band-width required is no greater with short than with long waves, it follows that if similar selective methods of reception are to be used and the available band of frequencies fully exploited, the permissible frequency-variation (expressed as a percentage) becomes much smaller. The frequency of the ideal transmitter would conform to four quite distinct requirements :

(1) It would be correct.

(2) It would not "drift," that is, change slowly with time.

(3) It would not "scintillate," that is, change from instant to instant.

(4) It would be possible to alter it if occasion arises.

In order to obtain the first condition, accurate methods of absolute frequency measurement evidently are necessary, and these have now been so highly developed that measurements of frequencies can be made to less than one part in a million.

With regard to the second condition, several different types of constant frequency sources are now available and will be discussed in this chapter. Such circuits produce only a very small output and a chain of amplifiers is required before sufficient power is available to drive even a small transmitter. It has already been pointed out (Chapter X) that a self-oscillator is but seldom used as a transmitter for short waves owing to its general frequency instability, and the main features of a driven system were discussed in Chapter X.

The third requirement is bound up with the design of the complete transmitter. Although a source of frequency having a very small drift may be provided, it by no means follows that the frequency radiated from the transmitter will be perfectly steady. It may be found to vary at a rapid rate as soon as the transmitter is modulated or keyed. This is because the changing loads on the transmitter are reacting back on to the frequency source in some way. To prevent this the driving source should be kept oscillating and drive into a small but constant load at all times. This is best accomplished by arranging the first amplifying stage, the isolator, to be unaffected by subsequent modulated stages, and if possible operating the valve of this amplifier without grid current.

Further, if the master oscillator be run at a lower frequency than the main power amplifier, this in itself will tend to prevent reaction effects. There appears to be an idea prevalent that it is desirable to produce master-oscillator circuits at the same frequency as the final frequency to be radiated. Actually it is a considerable advantage to employ one at a lower frequency, not only because of immunity from reaction effects, but because a larger choice of final frequencies becomes possible from one master oscillator. The fourth requirement is a difficult one to cater for, but in some cases a necessary one to provide and may prove the deciding factor in choice of master oscillator to be adopted.

It is now possible commercially to produce transmitters the frequencies of which do not drift by more than 1 part in 10^6 over very long periods (although a tolerance of 1 part in 20,000 is sufficient for most purposes) and which scintillate by an amount too small to measure.

In the case of short wave telephony it is essential to reduce the scintillation to very small proportions, apart altogether from the question of interference with other channels, otherwise bad distortion due to multiple echoes is produced. We have, therefore, to consider suitable types of circuit to generate a very constant frequency and capable of delivering sufficient output to drive a chain of amplifiers. Before the need for such great precision arose, earlier driven systems were controlled by a fairly large (100 watt) drive, consisting of an ordinary valve self-oscillator, working at the frequency to be radiated,

but it was found difficult to design a stage of this size to be free from frequency drift, particularly on first starting up.

As the permissible tolerance became less, wireless engineers looked around for means of obtaining a more stable driving source, and three distinct types of "constant frequency" drives have been evolved, all of which are maintained in oscillation by valves, but the resonating element may be :

(1) Tuning fork or steel rod.
(2) Piezo-electric crystal.
(3) Special electrical-resonant circuit.

The valve-maintained tuning fork, due to Eccles, was the first type developed for long waves, and since this is of necessity a low-frequency source very many frequency-multiplying stages are necessary before it can be of use to control a short wave circuit.

The second type, developed by Cady and Pierce as a result of original work by the Curie brothers, makes use of the piezo-electric properties of quartz, tourmaline or Rochelle salt. Such crystals can operate efficiently up to 15 megacycles, and thus but few frequency-multiplying stages are necessary even for short and ultra-short wave circuits.

The use of mechanical oscillators temporarily eclipsed the electrical resonant-circuit oscillator, largely on account of the effectively high Q value which is obtained by such means and the ease with which a circuit of zero temperature-coefficient can be obtained, but recent intensive work on the technique of the electrical resonant circuit oscillator has established it again in the forefront, and for certain purposes it compares favourably with mechanical drive systems.

The choice of what type of master oscillator to use is often difficult. All types of constant-frequency, master-oscillators are expensive, the main item not being that of the actual components used but the cost of final setting up and calibration, this rising rapidly as the tolerance is reduced.

A tuning fork would not now be considered as a source of frequency for a radio transmitter, as the quartz crystal is more stable and produces a frequency having a much more convenient value. Where very great precision is required and the frequency can be decided without doubt, the crystal will usually

be chosen, but if there is any question of flexibility, then the electrical resonant circuit will be preferred.

An oscillator will be a part of most radio receivers, since the superheterodyne type is the most common. Such oscillators need to be as stable in frequency as possible but must usually be simple in design, small in size and capable of tuning over a wide range.

In this chapter the general principle of oscillators depending upon feed-back will first be discussed. Then some circuits will be given, including those useful at the highest frequencies for which feed-back oscillators are suitable. The factors upon which frequency stability depends will next be considered, followed by a description of the quartz oscillator. Finally, means of measuring radio frequencies will receive attention.

The Oscillator as an Amplifier with Feed-back

Consider the circuit shown in Fig. 245, consisting of an amplifier having a gain μ, back-coupled through an attenuating network, so that βE_o is fed back into the input. Both μ and β will, in general, be complex quantities, since there may be a phase shift in both the amplifier and the reducing network.

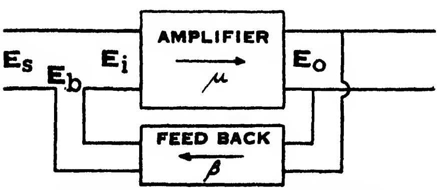


FIG. 245. Illustrating Feed-back.

We have

$$E_o = \mu E_i$$
$$E_i = E_s + E_b = E_s + \beta E_o$$

Hence

$$E_o = \mu E_s + \mu\beta E_o$$
$$E_o(1 - \mu\beta) = \mu E_s.$$

The gain of the amplifier with feed-back is, therefore,

$$\frac{E_o}{E_i} = \frac{\mu}{1 - \mu\beta}.$$

It will be seen that it is possible for $\mu\beta$ to equal unity, when the smallest input would give an output, that is, the arrangement has become a self-oscillator.

This discussion is quite general, any type of amplifier being capable of oscillating in this way. Mechanical, acoustical or

electrical systems, or combinations of these, are possible. An example of a mechanical amplifier set to self-oscillation is the reciprocating steam engine, a small portion of the mechanical output being used to drive the valve mechanism and so control the available energy in the steam.

An example of an acoustic-electric self-oscillator is the howl of a public-address system when the gain is increased too much.

It would appear that when $\mu\beta = 1$, the oscillations would become infinite, but this supposes that μ is a constant irrespective of the voltage being handled. Actually, the value of μ will decrease when E_i exceeds a certain value. In the case of a valve, for example, when the grid voltage swing is so great that the top and bottom bends of the characteristics are in use, the effective value of μ over a cycle will be reduced. The steady amplitude of oscillation will, therefore, be that for which a balance is obtained between the effective amplification and the feed-back, so that the gain round the whole circuit is unity.

Since E_i is to be provided entirely by E_b when the arrangement is oscillating, it follows that the total phase shift around the circuit must be zero, 360°, or a multiple thereof. This will decide the frequency of oscillation.

A valve stage usually produces approximately a phase reversal and therefore the feed-back circuit must also produce a 180° phase-shift, approximately. It is not necessary to have a resonant circuit to determine the frequency and some low-frequency oscillators employ resistance-capacity networks arranged to produce 180° phase shift at the required frequency. Circuits employing resonance (either electrical or mechanical) are always employed at radio frequencies, however, as they are more stable.

Types of Oscillators with Electrical Resonant Circuits

If the circuits of Figs. 246 and 247 are examined, it will be seen that it is possible for the voltage applied to the grid to be

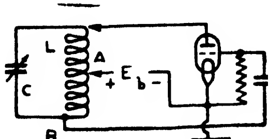


Fig. 246. Hartley Oscillator.

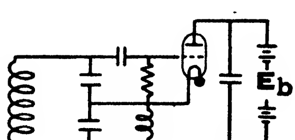


Fig. 247. Colpitts Oscillator.

opposite to that in the anode circuit, and hence, with the phase reversal in the valve, the total phase shift can be zero, as required for oscillation.

We have already seen that the amplitude of oscillation will be determined by the curvature of the valve characteristics. Conditions will be similar to those in a Class C amplifier and the anode current will be far from sinusoidal. We are therefore not justified in drawing vector diagrams of valve oscillators,

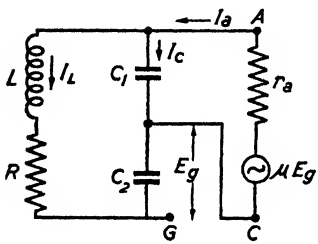


FIG. 248. Equivalent Circuit of Colpitts Oscillator.

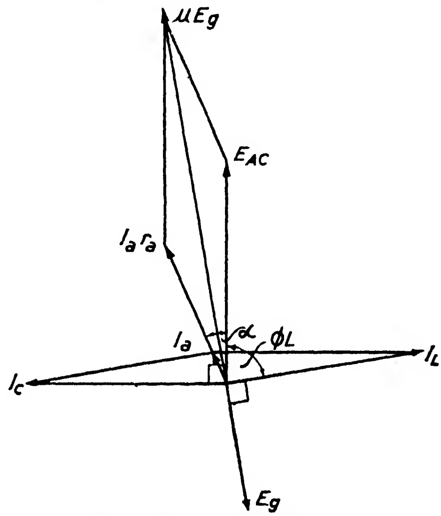


FIG. 249. Vector Diagram for Colpitts Oscillator.

except that these can tell us whether the conditions are right for oscillations to build up and will give us an approximate idea of what is happening in the circuit.

Let us construct a vector diagram for the Colpitts circuit. Consider the condensers to be free from loss but allow for a resistance in the coil (which resistance would also take into account any power drawn from the oscillator). The equivalent circuit for the oscillator is then as shown in Fig. 248. If we draw the vector E_{AC} (Fig. 249) for the voltage between anode and cathode of the valve, I_c must lead 90° on this. I_L will be given by

$$\frac{E_{AC}}{\sqrt{R^2 + \left(\omega L - \frac{1}{\omega C_2}\right)}}$$

and will have a phase angle ϕ_L where $\tan \phi_L = \dfrac{\omega L - \dfrac{1}{\omega C_2}}{R}$.

The voltage E_g applied to the grid will lag 90° behind I_c. This gives us the direction and magnitude of μE_g, which must be the vector sum of E_a and $I_a r_a$.

It will be seen, therefore, that there is usually a phase angle between E_{AC} and I_a, that is, the frequency of oscillation is not exactly the frequency of parallel resonance. It could only become so if the losses in the coil were negligible, or if the condensers also had a phase angle of suitable value. Since I_a is leading on E_{AC}, it follows that the oscillation frequency is higher than the resonant frequency of the circuit.

When such circuits are used at the higher radio frequencies

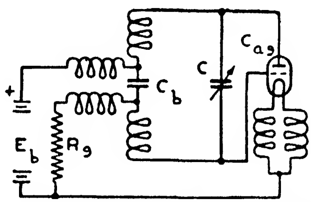
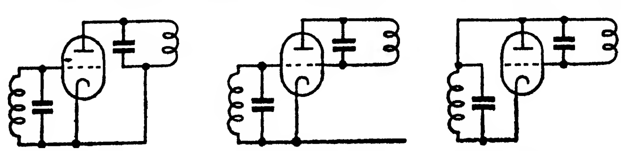


FIG. 250. Franklin Short-wave Oscillator Circuit.

the anode-cathode and anode-grid capacitances have a sufficiently low reactance that they provide considerable feed-back and modify the performance of the oscillator. This feed-back may, in fact, be all that is required and hence the circuit shown in Fig. 250 (due to Franklin) is frequently employed for short waves. The circuit diagram appears to be very similar to



(a) Common Cathode (b) Common Grid (c) Common Anode

FIG. 251. Types of H.F. Oscillators.

that of the Hartley, but in practice the coils are usually completely uncoupled and feed-back is through the valve capacitance.

At frequencies high enough to make the capacity reactance

between the valve electrodes sufficiently low, the circuits of Fig. 251 can be used. The D.C. circuits are not shown on these diagrams. It will be seen that each oscillator comprises two tunable circuits, one electrode being common to two circuits. The common-cathode circuit is usually referred to as the tuned-anode, tuned-grid oscillator.

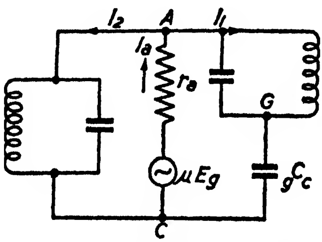


FIG. 252. Equivalent Circuit of Common-anode Oscillator.

It can be shown that each of these circuits is capable of oscillation for a certain set of adjustments. In the case of the tuned-anode, tuned-grid oscillator, if the resonant frequencies of the two circuits differ appreciably, then the circuit which is tuned to the lower frequency will be the

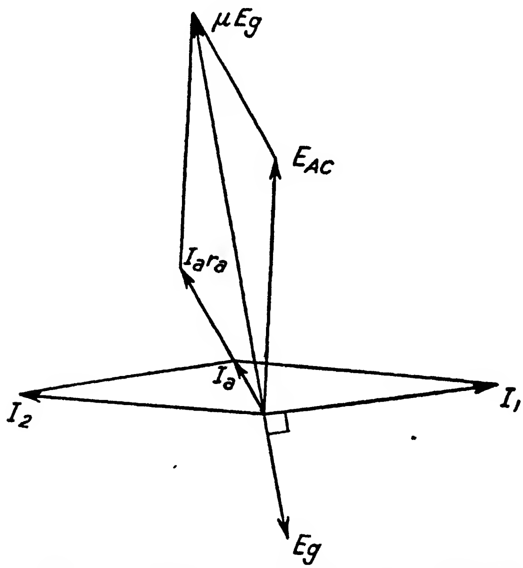


FIG. 253. Vector Diagram of Common-anode Oscillator.

controlling one, the tuning of the other having only a small effect. To get a good output, the anode circuit should be the controlling one, as oscillations will then take place at very nearly its resonant frequency and the load will therefore be largely resistive.

We will now construct a vector diagram for the common-anode arrangement (Fig. 251c). Electrode capacitances between anode and cathode and between anode and grid are in parallel with capacitances in resonant circuits and can therefore be included in them. The grid-cathode capacitances will, however, be the path through which feed-back is obtained. The equivalent circuit will be as in Fig. 252. If the anode-grid circuit is tuned to a frequency slightly higher than the oscillation frequency, then it will be a large inductive reactance. The grid-cathode capacitance will not have a large reactance at the very high frequencies for which this circuit is used, and hence I_1 lags nearly 90° behind E_{AC}. E_g must lag 90° behind

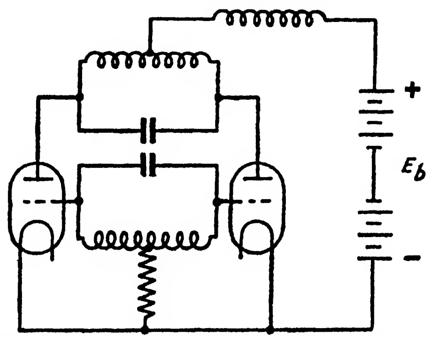


Fig. 254. Push-Pull, Tuned-Anode, Tuned-Grid Oscillator.

I_1 and this gives us the position of μE_g. It follows that I_2 must lead on E_{AC}, that is, the anode-cathode circuit must be tuned to a lower frequency than the oscillation frequency.

The oscillator will normally be adjusted so that the anode-grid circuit is more nearly in tune to the oscillation frequency than the anode-cathode circuit. Under these conditions the anode-grid circuit is the main frequency control and the tuning of the other mainly decides the amount of feed-back. With all these circuits (including the so-called tuned-anode, tuned-grid oscillator) it is most important to observe that *both* circuits must not be tuned to the same operating frequency. In such a case the anode current becomes so misplaced with the grid driving voltage that highly inefficient operating conditions result.

The circuits (a) and (b) of Fig. 251 can be studied in the same way.

Push-pull arrangements are very suitable and Fig. 254 shows a circuit very similar to that of an amplifier but without the neutralising condensers. The action of this circuit is similar to that of the tuned-anode, tuned-grid oscillator. It may sometimes be necessary to fit the neutralising condensers and to partially neutralise, as otherwise the feed-back through the valve capacitance may be too great.

Oscillators for Ultra-High Frequencies

The complications which arise when we attempt to use valves at very high frequencies have already been discussed in the previous Chapter. We have seen that the lead inductances and capacitances greatly modify the valve performance. Hence, if we set up an oscillator circuit which appears on paper to be the same as a low-frequency circuit, it may behave quite differently.

Valves for use at ultra-high frequencies will have the leads as short and well-spaced as possible, but it is not possible to reduce the cathode inductance very much. This is for the practical reason that the cathode is at a high temperature, whilst the seal through the valve envelope must be at a much lower one. The only simple way to ensure this is to have a fair length of lead between cathode and seal.

Most ultra-high frequency oscillators employ two tunable circuits, because arrangements such as the Hartley or Colpitts are so modified in performance by valve reactances as to prove difficult. Of the three alternatives shown in Fig. 251, the common-cathode circuit is normally chosen for lower-frequency use, because it is satisfactory and convenient. At ultra-high frequencies, however, it is not the best arrangement, partly because of the effects of cathode-lead inductance and partly because, for the same values of valve capacitances, the other circuits will tune to higher frequencies. The circuits of either Fig. 251b or 251c will, therefore, be chosen.

The capacitances to earth will be important and it will usually be preferable to earth one electrode. Here, again, we depart from the usual low-frequency practice of earthing the cathode, and earth either anode or grid. It is often convenient to earth the common electrode but this is not always done. In a power oscillator, a convenient arrangement is to earth the anode, particularly if the valve is of the metal-glass type and

the anode is external. The anode can then be fitted with cooling fins and, owing to its bulk, will have the largest capacitance to earth of any of the electrodes.

It will be desirable to screen the circuits one from the other, so that the valve capacitance forms the feed-back circuit.

Push-pull circuits are much used, in order to get more power output and also because of their symmetry. In the case of all ultra-high frequency oscillators, it will be necessary to correlate closely the design of the valves and the layout of the circuits in which they are to be used. Suitable forms for the resonant circuit will next be discussed.

The ordinary coil and condenser circuit becomes very small when it is to tune to frequencies above 100 Mc/s. Also, self-capacity of the coil and inductance of the condenser leads and plates become so important that it is not even approximately true that the current is the same in all parts of the circuit. At these frequencies it is better to use circuits in which L and C are distributed according to a known law. The size then becomes more convenient and the Q can be much higher.

One such example is by the use of a resonant line, having properties which we have already discussed in Chapter VI. By suitable design this can have a Q much larger than the ordinary LC circuit, especially at high frequencies. For such frequencies it can be made very rigid mechanically, so as to hold the oscillator frequency as constant as possible. The valve electrode capacitances will be connected across the line and will, of course, decrease its length for a given frequency. The effect will be much less, however, than if the same capacitances were connected across an LC circuit.

An example will demonstrate this. Suppose that the capacitance between grid and anode of a valve is $5\mu\mu$F, whilst the inductance of a straight lead connected between them is $0{\cdot}1\mu$H. Then this, the smallest possible resonant circuit, has a frequency of about 225 Mc/s.

If we use a short-circuited line as a resonant circuit, then this must provide an inductive reactance equal to the capacity reactance of the condenser, so that the whole may behave like a parallel resonant circuit. Hence $Z_o \tan \theta = \dfrac{1}{\omega C}$ where θ is the electrical length of the line.

If Z_0 is 500Ω, then, from the above relationship, the length of line for 225 Mc/s would be about 6 cm and this could probably be reduced, thus increasing the possible oscillation frequency. Moreover, if the oscillator contains two line circuits, as is usual, then one of them can be much longer if this will simplify the design. This is because a $\frac{3\lambda}{4}$ line will show the same input impedance as a $\frac{\lambda}{4}$ line. Thus a frequency of 675 Mc/s could be obtained by using a line about 24 cm long, if line and valve capacitance were equivalent to a $\frac{3\lambda}{4}$ line.

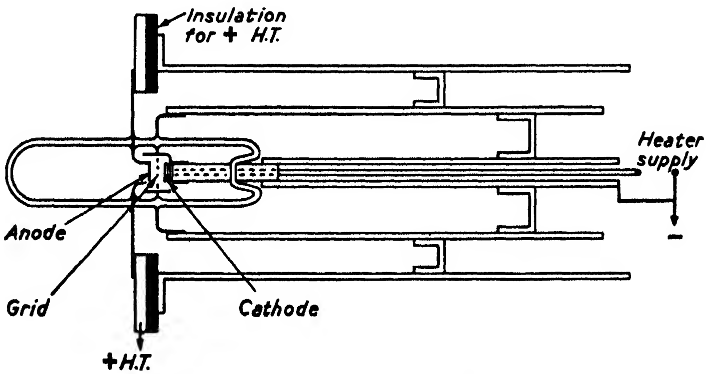


FIG. 255. Line Oscillator.

Both lines cannot be treated in this way, however, because then the oscillator would work at the lower frequency for which the lines were approximately $\frac{\lambda}{4}$ long.

We see also that if Z_0 is decreased, then the effect of the valve capacitance is reduced. If a concentric line of 80Ω characteristic impedance were used, for example, then its length would be 22·3 cm in order to tune as a $\frac{\lambda}{4}$ line (including the 5$\mu\mu$F valve capacitance) to 225 Mc/s.

Two typical oscillators using lines will now be described. A single-valve oscillator which will give 5 watts at several hundred megacycles and will give 0·5 watts at 3,000 Mc/s is shown in Fig. 255. This is of the common-grid, earthed-anode type.

In the valve used the electrodes are planar, not cylindrical, and both anode and grid are brought out through disc seals running right round the valve. The anode-cathode capacitance (which might otherwise be insufficient) is increased by leads attached to the cathode, which pass through holes in the grid disc. The anode is connected externally to a fairly massive metal structure and this helps to dissipate the heat.

It will be seen that each tuned circuit is formed by a con-

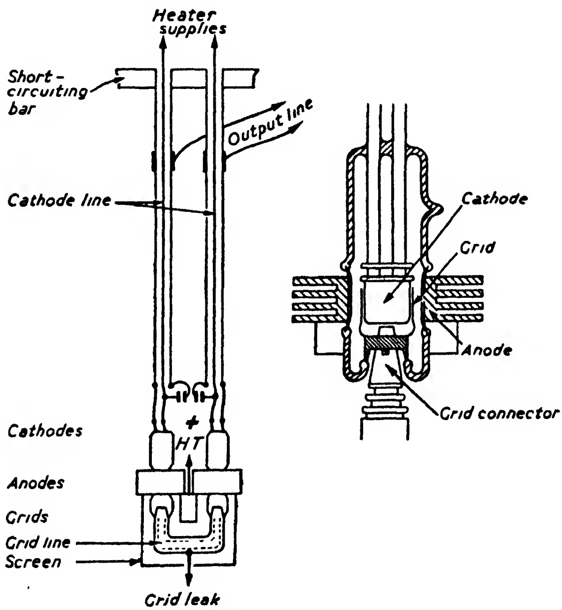


Fig. 256. Push-Pull Line Oscillator.

centric line and that they are one inside the other, so that the earthed outer conductor of the anode-grid circuit surrounds all. The valve has an indirectly-heated cathode but one side is connected, within the valve, to the cathode. The other heater connection runs inside the inner conductor of the grid-cathode line.

A push-pull circuit of the common-anode, earthed-anode type is shown in Fig. 256. This was used as a radar transmitter on about 600 Mc/s developing a peak power of 100 kW when the pulse duration was $2\mu s$ and the repetition frequency

600 c/s, so that the mean power was 120 W. In the valves used (CV92), the anodes are external and have cooling fins. The grid comes out at the bottom in a massive seal (to reduce inductance and enable large currents to be carried) whilst the cathode connections come out through the top.

The cathode is of the oxide-coated type. The use of oxide coatings in conjunction with high anode voltages (in this case 8 to 10 kV) and high emission currents was developed during the war in order to provide the large emission necessary for the high peak powers of radar transmitters. The very small grid-cathode clearance will be noticed from the diagram.

It is necessary for the anode-grid circuit to be slightly less than $\frac{\lambda}{4}$ and this can only be managed by having a screen round the massive grid line, as shown. The cathode circuit can then be operated at somewhat more than the $\frac{3\lambda}{4}$ adjustment, so that there is no difficulty in this. The output is taken off on the cathode circuit at a suitable point to give impedance matching.

The circuit shown is for fixed-frequency operation, but a modification has been much used in which a sliding plate varies the tuning of the grid circuit, giving a range from about 500 to 600 Mc/s.

Frequency Stability of Oscillators with LC Circuits

A great deal of work has been done on this subject, notably by Franklin, Witt, Thomas and Lea. Mathematical investigations have proved intractable, owing to the non-linear characteristics of an oscillator, and progress has been made by careful qualitative analysis of the problem and much patient experiment. The oscillating frequency is primarily determined by the product *LC* and we must therefore arrange to keep this as constant as possible. The factors producing the largest changes are temperature and mechanical vibration, though atmospheric pressure, humidity and other effects also play their part.

The resonant frequency of an *LC* circuit, in which no precautions have been taken, may vary as much as 1 part in 1,000 for quite ordinary temperature changes, and where vibration is

considerable, as in ships and aircraft, carefully designed suspensions may become necessary for units containing oscillators.

In oscillators claiming to be stable sources of constant frequency it will be necessary to design coils and condensers so that their variation of L and C with temperature is consistent, and preferably small, and then fit a thermal-compensating component, usually a condenser. This will be adjusted experimentally to make the frequency variation with temperature of the whole oscillator as small as possible.

Coils which show a small variation of inductance with temperature have been made by depositing the conductor on to a former of suitable material. In this way the expansion of the former becomes the controlling factor and some suitable ceramic materials have much smaller coefficients of expansion than metals. These coils are not so stable as might be expected, however. It appears that, on the edges of each deposited turn of the winding, there are particles of metal which are not properly in permanent contact with the turn. If these make contact occasionally, then the constants of the coil, particularly its self-capacity, change. These coils have also the disadvantage that only small inductances can be obtained with coils of reasonable size.

For general use a single-layer winding on a grooved, ceramic former is suitable and the winding is put on so that it is just not tight at the lowest temperature at which it is required to work. The radial expansion will then be definitely governed by the coefficient of expansion of the wire.

Coils will need to be screened and the spacing between coil and screening can should be ample, otherwise any change of temperature which alters the dimensions of the can will change the inductance appreciably.

Coils have been designed in which the effect of axial expansion (which reduces the inductance because it reduces the mutual inductance between adjacent turns) balances the effect of radial expansion (which increases the inductance).

It will be necessary to split the circuit capacitance between a fixed and variable condenser, although this, of course, reduces the frequency range which can be covered. The variable condenser can then frequently be of an ordinary, well-made type, which will have a repeatable frequency temperature

coefficient. That is, if the temperature of the condenser is raised and lowered a number of times, it will be found that the same frequency changes occur. It will be very important to have good bearings for the condenser, because axial movement between fixed and moving vanes makes such a large difference to the capacitance.

Considering the fixed condenser, some of the condensers employing electrodes deposited on a ceramic of high permittivity have a very small temperature coefficient and are very small in size. It is found, however, that these condensers may, at times show a sudden change of capacitance, due it is believed, to the same causes as the erratic changes sometimes occurring in coils made of metal deposited on ceramic.

Air-dielectric, fixed condensers are preferable where space will allow. These can be designed by choice of materials and shape to have a very small and consistent temperature coefficient.

One type of thermal compensating condenser is described on page 430.

The variation of oscillator frequency with changes of humidity has not received the same attention as temperature changes, but has been shown by Lea to be of considerable importance. A change of 1 in 1,000 occurred in the frequency of a certain oscillator when the humidity was changed from 50% to 90%. The change is due to change of condenser capacitance, and this, in turn, is due to two causes—the change in dielectric constant of the air between plates and a moisture film on the plates.

Variation of frequency with pressure is likely to be very small over the range of barometric changes usually encountered.

Turning now to variations of frequency caused by other components than the coil and condenser, it will be evident that the maintaining-valve, electrode-capacitances are in parallel with the *LC* circuit and therefore help to determine its resonant frequency. The effective value of these capacitances, when the valve is working, depends upon the operating point on the valve characteristics and hence upon the supply voltages. It will, therefore, be desirable to stabilise the anode supply of an oscillator, when a constant frequency is required. A change of valve may produce a large change in calibration.

If we look back at the vector diagram for the Colpitts

oscillator, as an example, we see that the frequency generated is not usually exactly that of the resonant circuit. Suppose that the effective r_a changes, due to either a change of anode supply volts or a change in amplitude of oscillations, then α will have to change and this means a change of frequency. If the resonant circuit has a high Q, then the necessary phase shift will be brought about by a very small change of frequency.

The desirable high Q should not be obtained, however, by a high L/C ratio, because this will enhance the effect of changes in valve capacitances.

The effective Q of the circuit includes, of course, the effect of any load and it is therefore desirable that only a very small (and constant) amount of power should be drawn from the oscillator. Any change of load will vary α and hence the frequency.

Since the oscillator will normally be working over the curvature of the valve characteristics, a change in the amplitude of oscillation will change the waveform. It can be shown that if the amplitudes of the harmonics change, then the oscillation frequency will have to change in order to preserve the necessary phase relationships.

When considering changes of frequency due to variations of temperature, we had in mind slow alterations in ambient temperature after the oscillator had been switched on for some time. A period will elapse after switching on, however, during which there may be a large "drift" of frequency due to components—particularly the valve and any other components near it—warming up as they come into operation. It may be necessary to refrain from using the oscillator during the warming-up period. Transmitters working on services for which regulations prescribe a close frequency-tolerance, will usually need to keep their master-oscillators running continuously. Transmitters which are allowed a wider tolerance and for which it is not feasible to keep the oscillator running all the time—such as small, marine transmitters—should have oscillators for which the warming-up time is very small or compensation should be provided to cover the warming-up.

Looking back over the requirements for a constant-frequency oscillator, we can realise that they are difficult to obtain except at very low power. Further, their output is delivered to a

"buffer" or isolator valve, run without grid current, so that the load on the oscillator is very small and constant. This isolator is often a frequency multiplier, which still further reduces the reaction back to the oscillator, of any succeeding amplifier stages.

We will now describe one constant-frequency valve oscillator which has been much used for driving transmitters, so that one method for satisfying the requirements may be studied.

Franklin Master-Oscillator

An interesting type of precision oscillator is that developed by Franklin and Witt, the circuit diagram of which is shown

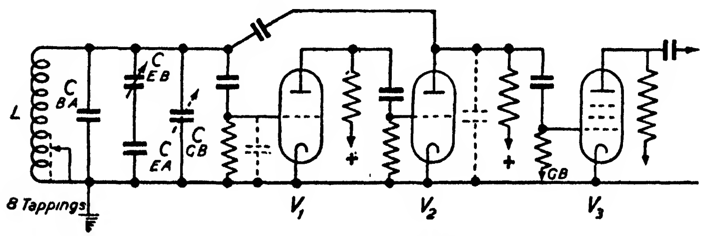


FIG. 257. Circuit of Franklin, Master Oscillator.

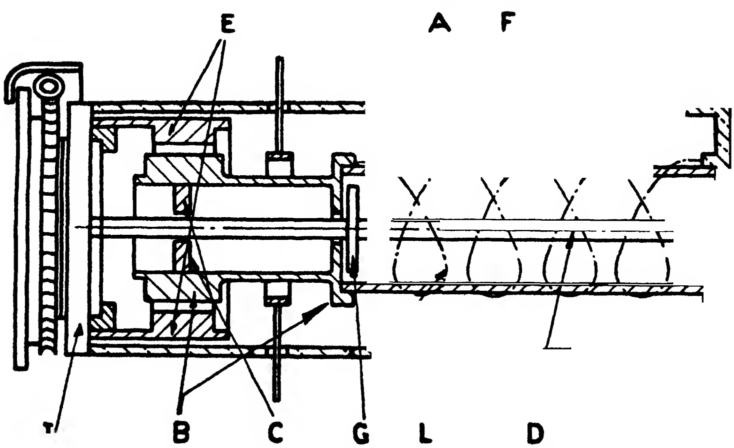


FIG. 258. Construction of Franklin, Master Oscillator.

in Fig. 257, and the mechanical arrangement in Fig. 258.

Associated with the resonant circuit are two valves V_1 and V_2. Two valves are employed partly because the higher gain

thereby obtained enables the maintaining circuit to be coupled very loosely to the resonant circuit and partly because one end of the resonant circuit may then be earthed, which makes the mechanical design easier. Since the voltage applied to the grid of V_2 is opposite in phase to that applied to V_1, it is evident that the anode of V_2 should be coupled back to the same end of LC as the grid of valve V_1. The coupling capacities are only about 1$\mu\mu$F and therefore smaller than the valve capacities, and hence the effect is that of a circuit having low grid and anode taps.

The oscillator is arranged to work at a lower frequency than that finally required, a usual and desirable feature of most constant-frequency drives, and the valve V_3 is followed by a normal frequency-multiplying amplifier.

The usual type has eight ranges, obtained by the tappings shown on the coil. The total range of frequencies covered is about 1 : 1·35. The cylindrical, brass case A, which is earthed, forms one side of the circuit and acts as a support for the unit as a whole. Concentric with the cylinder is the coil former, of special ebonite. This carries the inductance winding L, one end of which is connected to the end of the case and the other to the insulated condenser plate B. This is of half-tubular form and is built on the free end of the coil former, its outer end being mounted on the insulating diaphragm C, which rides on the brass rod D.

Between the condenser plate B and the case is a second pair of circular half-plates EE, insulated from both and carried on an insulating end-cheek J. The circular half-plates EE can be rotated by means of a worm wheel and this constitutes the tuning control, sufficient variation being provided to give about 10% variation of frequency. It will be seen that the variable capacitance between EE and B is in series with the fixed capacitance between EE and A. Also, there is a direct capacitance between B and the coil to A. The circuit can therefore be represented by the diagram of Fig. 257.

Thermal compensation is carried out by the condenser formed by G and the end of B. The gap between these is varied by rotating the screwed rod D and this will decide the amount of compensation.

The operation is as follows :

By proper selection of materials the chief endwise expansion is designed to take place in the insulating former and thus, as it warms up, the plate *B* moves away from the compensating plate *G*. The axial expansion of the winding causes its inductance to be increased, but this increase is offset by the reduction of capacitance between *G* and *B*.

To make the drive immune from external capacity changes and vibration, and to some extent to prevent too sudden a temperature change, the whole unit is slipped inside a second brass cylinder (not shown), between the two being a layer of thick felt. This second cylinder forms a support for the two oscillator valves and the coupling valve, and thus a self-contained compact unit results. The coupling condensers, which of course must be connected to the free end of the coil, are simply two very minute brass strips held in a mica ring, connections being made by two pins projecting through both cases.

Tests on these oscillators show that a rise in temperature from 25° to 45° C. resulted in a permanent frequency change of about 1 in 10^5. The oscillator takes about six hours to settle down completely to the new temperature, however, and during this period the frequency may vary by 5 in 10^5. The change in frequency for a 1% change of supply volts is about 2 in 10^6.

We now propose to discuss the quartz oscillator and will commence by studying the properties of quartz itself.

Quartz

Certain crystalline substances, notably quartz, tourmaline and rochelle salt, exhibit a phenomenon known as the piezo-electric (or pressure-electric) effect. Of these, rochelle salt, although the most active, is unstable physically and it is not greatly employed. Tourmaline is used occasionally, but for most wireless purposes quartz is preferred and we shall only consider this. Quartz is a very common form of silica, but good crystals are only found in quantity in Brazil, Madagascar and Japan, and as only a very small percentage of the quartz found is of any value for piezo-electric work, very great care is required in the selection of the raw material if large wastage is to be avoided.

A natural quartz crystal can appear in various forms but

an idealised crystal should take the form of a hexagonal prism terminating at both ends in a pyramid. The quartz found, however, is far from perfect and in many cases badly deformed with faces of uneven size and one end larger than the other.

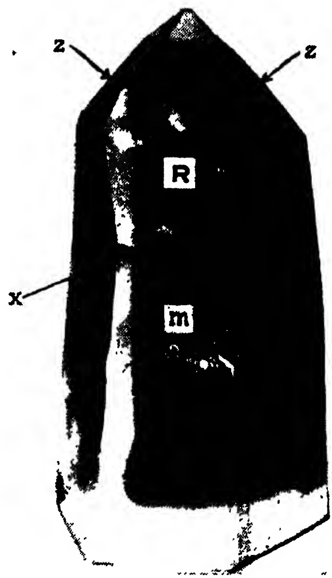


FIG. 259. Typical Sample of Natural Quartz.

Generally speaking, specimens are found broken and a natural crystal will usually taper as shown in Fig. 259.

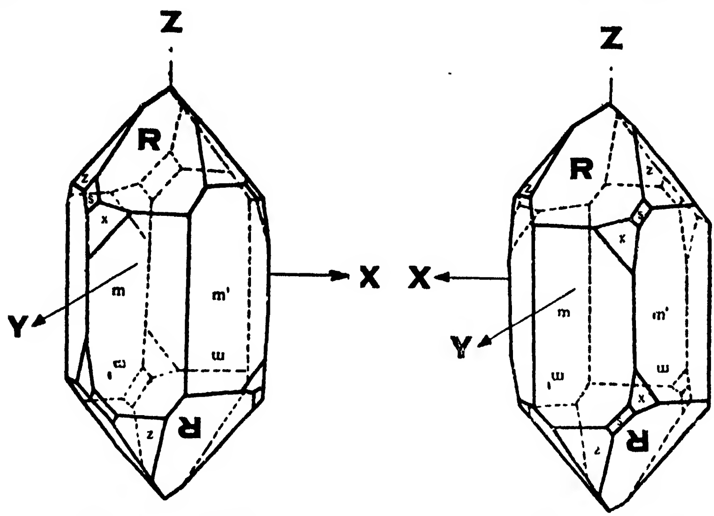
Nomenclature

Much confusion arises in dealing with the nomenclature of quartz, as not only have crystallographers not been consistent, but various investigators in piezo-electric work have each produced their own particular system of notation which appeared to be best suited to their needs. This means that the reading of current literature is difficult. In an attempt to get rid of the confusion arising, a notation was suggested by the National Physical Laboratory, but it has not been widely adopted, possibly because its manner of presentation is more suited to the scientist than to the practical works engineer.

In the sections which follow we propose to adhere largely to the suggested N.P.L. notation but to correlate it, where it appears desirable to do so, with notation adopted by other workers.

Types of Crystal

Referring to Fig. 260, which shows idealised types of double-ended crystals, the main body contains three pairs of parallel faces designated *m*, *m'*, capped at each end by a hexagonal pyramid, the six faces of which are denoted alternately by the



Left-handed Crystal. Right-handed Crystal.

FIG. 260.

letters *R* and *z*. The position of the crowns at the opposite ends of the crystal is such that the *R* (and *z*) faces are above one another, but the crown is skewed so that the *R* face of one is directly above the *z* face of the other crown. We may really regard the *R* and *z* faces of the crowns without the body as forming the sides of two interleaved rhombohedrons. Although the size of the various faces varies considerably, their angular relationship is very definite, however distorted the crystals may be, and this angular relationship can act as a guide when making cuts. Thus the included angle between adjacent *m m* faces on the basal plane is 120°, and the angle between an *R* face and the optical axis is 41° 47′.

In certain crystals, facets *s* and *x* occur at the junction of the pyramid as shown in Figs. 259 and 260, and the order of these facets gives the clue to the form of crystal. The surface of a crystal usually shows growth lines across the *mm′* faces, seen in Fig. 259, which in a good crystal are parallel and unbroken.

Quartz may assume one of two simple forms which are the mirror image of each other as seen in Fig. 260; or a complex crystalline structure may result, known as twinning, the latter type of crystal being of no value for piezo-electric work.

Examples of perfect twin crystals are shown in Fig. 261, but it should be explained that internal twinning may be present in a crystal which to outward appearance is of simple form. The notation for crystal faces is shown in Table XVIII and the position of these faces can be seen by reference to Fig. 260.

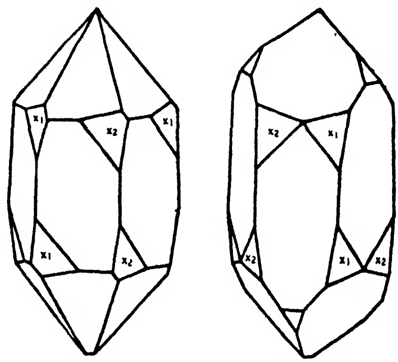


FIG. 261. Twinned Crystals.

TABLE XVIII

N.P.L.	R.C.A.
R	A
z	B
s	C
x	D
m (under *R* face)	E
m (under *z* face)	F

NOTE.—For convenience in discussion we propose to denote the *m* face under the *z* face as *m′*.

The simple forms, so-called right- and left-handed, both of which are equally suitable for piezo-electric work, are determined by whether the sequence of facets *m*, *x*, *s* and *z* follow the course of a right-hand or left-hand screw-thread. Or more simply, by whether the facets *s* and *x* are to the right or left

of an *R* and *m* face, when viewing the crystal with the *R* and *m* faces to the observer. Such an external feature indicates a molecular structure which will give a right-hand or left-hand

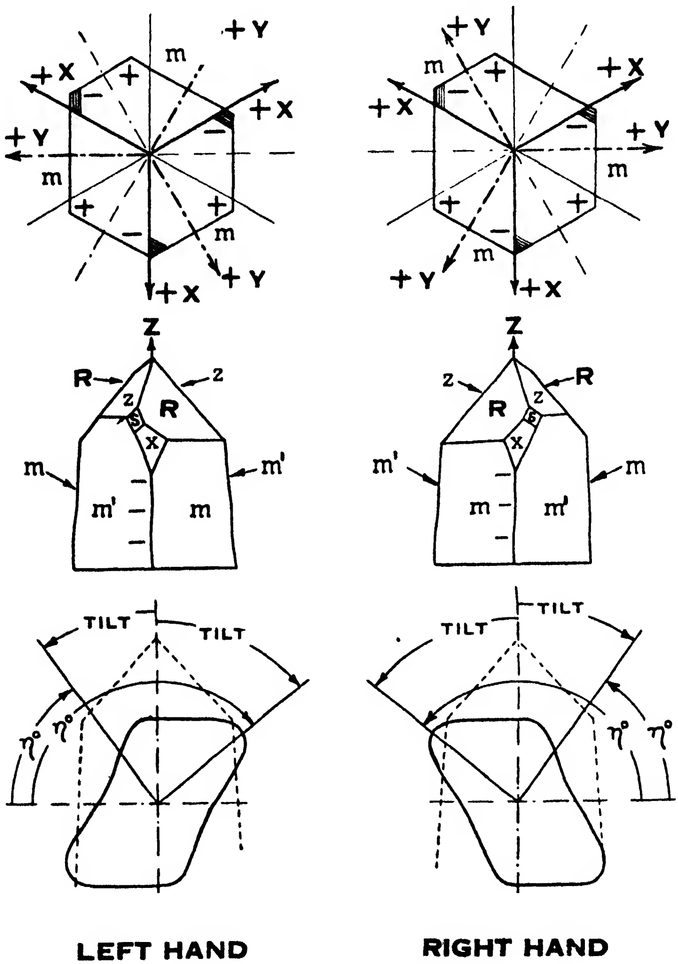


FIG. 262. Axes of Quartz Crystals.

rotation to the plane of polarised light transmitted along the optical or *Z* axis.

Thus if plane-polarised monochromatic light is passed through a slab of right-handed crystal cut normal to the optic

axis, the plane of polarisation is rotated in a clockwise direction when the observer looks at the source through the slab. It is also the direction in which a microscope analyser must be rotated by an observer who to restore light looks through the analyser towards the source. Amongst certain American workers a reversed convention is used but it will become evident later that, provided due precautions are taken, the " hand " of the crystal is quite immaterial.

The principal axes of the crystal are known as the Z, Y and X axes, which are all at right angles to each other. The Z or optical axis, which has already been mentioned, is that about which the rotary polarisation of light takes place. This axis

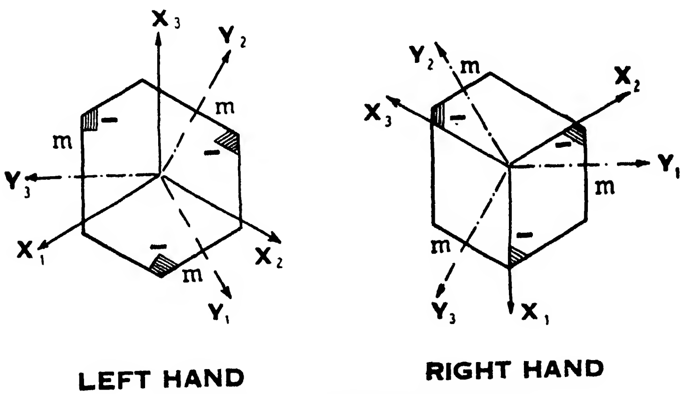


FIG. 263. Illustrating N.P.L. Notation.

is the same as the crystallographic axis of three-fold symmetry, because the structure repeats itself three times around it, and a knowledge of the axis is useful in the selection of suitable specimens and subsequent cutting.

The X or electrical axes, of which there are three, are each one parallel to an m face, and they have a two-fold symmetry. Thus in Figs. 262 and 263, the thick line axes, marked X, each at 120° to each other, are electrical axes having a positive sense. It is to be noted that only in a perfectly equal-sided hexagonal crystal will the X axes pass through the corners of the crystal.

At right angles to the electrical axes are the so-called mechanical axes, of which again there are three (shown chain

dotted and marked Y), each one of which emerges from an m face with positive sense, and is at right angles to an X axis.

With the N.P.L. notation, the rotational sense of a positive X axis to a positive Y axis is the same for both a left-hand and a right-hand crystal as shown in Fig. 263, but the usual mathematical notation is such that the rotational sense of the positive Y axis to the positive X, is clockwise for a right-handed, and anti-clockwise for a left-handed crystal, looked at from above, as shown in Fig. 262.

Actually, the point is not very important because, when orientating a crystal about the Z axis, the same results are obtained if it is orientated from a given axis (i.e. a point of symmetry), either clockwise, or anti-clockwise.

The Piezo-Electric Effect

When stressed by pressure or torsion, a quartz crystal may, in certain circumstances, exhibit a piezo-electric effect, that is, a mechanical force may produce an electric charge. Or conversely, when placed in an electric field it will exhibit strain and mechanical distortion.

Thus if a quartz crystal is compressed by pressure along one of its X axes, which implies an extension along a Y axis at right angles, the six edges of the crystal become charged alternately positive and negative, a negative charge always appearing at those edges carrying facets, or in the absence of facets, those edges which would have carried them had the crystal been so formed. The edges in which the compressed axis terminates exhibit the greatest charges ; thus in Fig. 262 (and 263), if pressure is applied across a pair of edges, say front to back (elevation), negative charges will appear on the three edges under the facets, shown shaded in plan, both with the left- and right-handed crystal. Such charges of course could be observed by means of a low-capacity electrometer.

It will be observed that when the position of the R and m face is determinable, even if no facets are present, it is possible to find the hand of the crystal by applying pressure along any X axis and noting the sense of the induced charges at the corners. As is seen from Fig. 262, if the induced negative charge is at the right-hand edge of an m face, the crystal is right-handed.

It is clear that the application of pressure along a Y face will have a similar effect (although it is not a usual procedure), except that the charges induced will be opposite in sign, but pressure applied in a direction along the Z axis produces no electric charge anywhere, neither does pressure applied evenly to all parts of the crystal (such as hydrostatic pressure), and the application of pressure to the crystal in directions other than X and Y will produce, as would be expected, complex, distributed charges.

As stated above, the converse effect also occurs, that is, the application of an electrical potential across an axis in the equatorial plane produces a contraction or extension along corresponding X and Y axes. Hence if we imagine a plate of quartz cut at an appropriate angle, subjected to an alternating electric field, it would be forced to expand and contract at the frequency applied. The mechanical oscillation will be of very small amplitude, however, unless the applied frequency is the same as the mechanical resonance of the plate which depends upon dimensions, the way in which it is cut, and upon the elastic constant of the quartz. We will see later that the frequencies of oscillation that can be produced are extremely high and thus we have a means of producing a convenient form of mechanical oscillator which can be maintained electrically at radio frequencies and piezo-electric control devices are becoming of increasing importance in wireless work. Before discussing the variety of ways in which crystal plates may be cut, it will be desirable to say something about the selection of suitable material.

Selection of Quartz

Since a large percentage of the natural quartz is useless for piezo-electric work, methods of preliminary inspection and test are of considerable importance. Although in some cases the external shape can give a clue to the goodness or otherwise of the crystal, it is often difficult even to distinguish the faces one from another on account of malformation. To give an example: out of 10 selected samples, with a possibility of 30 x and 30 s facets, only 5 x and 10 s were actually present. Further, quartz which to the eye appears a simple growth, as Fig. 259, may internally be malformed. Hence optical and

electrical tests become desirable before any cutting takes place.

A preliminary examination of the raw quartz is often made by placing it in a bath of winter green (Methyl-salicylate). Plane-polarised monochromatic light is projected along the Z axis, and the crystal viewed through an analyser along the Z axis, which will detect twinning. A further and standard test for the detection of one form of twinning is to polish and etch, with hydrofluoric acid, selected faces or sections and

FIG. 264. Quartz Slice under Polarised Light.

examine the etched surface pattern by eye.[11] Having eliminated crystals with twin growth, simple types exhibiting mechanical flaws are best avoided.

After selection of a specimen which is probably of good quality, the crystal will be sliced through normal to the Z axis and a specimen slice taken for optical examination using a polariscope.

A polariscope consists of a source of light, polarised by a Nicol prism, in line with which is a second Nicol, and a viewing eyepiece. The second Nicol is rotated so that light is extin-

guished, and if now the quartz section be placed between the Nicols, because of the optical properties of the quartz, light again passes in a manner which reveals the structure of the slab.

Suppose a perfect specimen of crystal, which has been cut exactly normal to the Z axis, to be placed in a polariscope in which a " white " light is used. Then the crystal would appear to be of uniform colour when viewed through the polariscope, the actual colour depending upon the thickness of the slab. When an average specimen is viewed, however, it will usually be found that only an area near the centre will show uniformity and the rest of the section will be broken up by a spectrum of colours whose brilliance and shape reveals the presence of imperfections in the main crystal. Fig. 264 shows a slice photographed through a polariscope, where the light parts represent the faults and are seen as a series of colour spectra. These areas are quite useless and will be marked so that they can be ignored in the subsequent cutting process. For the cutting of inclined plates (to be discussed later), it is essential to be able to identify the original positions of R and z faces with respect to the m faces. If this was not done by studying the pyramid top before cutting the specimen slice (possibly because of malformation), the various faces can be found by applying the electrical test mentioned on page 438, by viewing the slab through a polariscope, and by determining the hand of the crystal. This can be done by the use of plane-polarised white light viewed through an analyser and the employment of a single or double bi-quartz wedge. The electrical test determines the corners at which the s and x facets were above, and the hand of the crystal as determined by the polariscope and bi-quartz wedge determines whether the m face is to the right or left of the negatively charged corner as shown in Fig. 262. Thus although a knowledge of the hand of a crystal is of no importance in itself, it becomes necessary as a means of distinguishing between the m face below an R face and an m face below a z face.

As a result of the optical and electrical tests which have been indicated, it is possible not only to pick out the areas of bad quality from a crystal specimen, but also to identify the faces, and so obtain a datum from which to cut a crystal

to a specified angle, no matter how malformed the original specimen may have been.

Cutting

It can be imagined that with crystals of uneven shapes and weighing up to several pounds, the holding of a crystal so as to obtain a cut of specified angle to any given axis or face is a difficult matter, and up to now it is the usual practice not to attempt direct cutting, but to saw the block up into slabs cut normal to the Z axis and cut from such slabs as required. Cutting is carried out on what is virtually a milling machine, using for a cutter a type of lapidary's saw, that is a thin copper disc charged at the rim with diamond dust; or thin mild-steel discs are also used; or thin bakelite discs impregnated with carborundum or diamond. For sub-dividing into the finished crystal it is the practice to mount the rough blocks on to an accurately ground face-plate (often with thin ground-glass supports above and below) by an adhesive gum, and cut from the block the required crystal plates. A sufficient margin of material must be allowed for lapping to the finished size, the margin necessary being determined by the type of crystal-cut and the final precision required. The object of the glass supports is to prevent chipping of the crystal edges.

For the final finishing to size of quartz plates, various methods have been tried and are now in use. Hand lapping, the only successful method for many years, can only be carried out accurately by very skilled operatives and for the final process a stationary lap is employed. By such means accuracies to the order of ·005% are common. In terms of a 10 Mc/s crystal this would mean working to an accuracy of ·000013 mm. in a uniform layer, a surprising degree of precision. A mechanical lapping process has also been developed by placing the plates in a jar containing the final abrasive compound and slowly rotating the jar. Because the abrasive particles move slowly and lightly across the faces of the quartz, a very slow but linear change of frequency with time takes place and prediction of the final frequency is possible. Such a method, however, is only economic when a number of similar plates are required to be finished.

Alternative to the mechanical lapping methods of finishing

plates, there are a number of surface processes which are becoming more and more adopted in practice. The surface may be reduced in thickness by etching, usually with a weak solution of hydrofluoric acid ; or built up by the deposition of metal in close contact, either by sputtering, electrolytic, or evaporation processes, gold and silver deposits being usual. Such films can then be reduced, if necessary, by chemical means.

Although mechanical methods can give very good results it is found that plates so treated suffer from a drop in activity

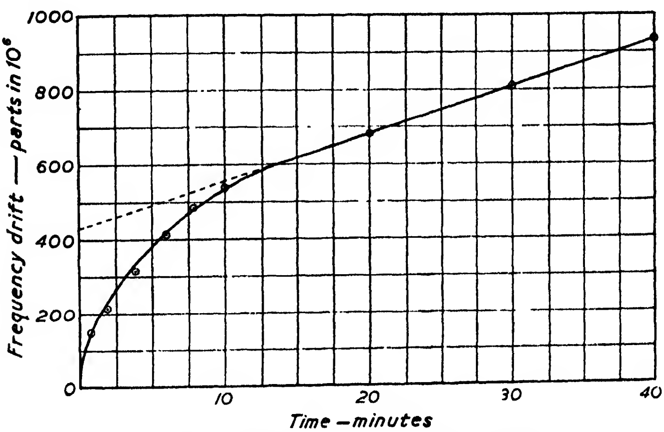


FIG. 265. Frequency Drift of Etched Plate.

and a frequency drift, known as ageing. Although the exact cause of ageing is yet to be fully explained, it is considered to be due to the change in the damping effects of the various extraneous material which gets packed into the surface layer of the plate during the grinding process. For this reason the finishing of quartz by an etching process, or combined etching and plating, is to be preferred, as the first effect of etching is to clean from the surface all occluded particles, thus ridding the crystal of this ageing effect. Thus Fig. 265 shows a frequency drift/time curve for an etched plate. It is observed that the curve, after 15 minutes, is linear with time. The intercept of this curve with the frequency-change axis is a measure of the loose quartz particles, abrasive and other foreign matter that was present in the plate after its final lapping, and which has been cleaned out by the etch. A some-

what similar effect can be obtained by the use of wet steam which is useful as a means of clearing a quartz plate of occluded materials rapidly.

Types of Cut of Crystal Slices

We propose to discuss, first, plates of the simple so-called Z, X and Y cuts, whose main definition is generally agreed by all writers, and to which all other types of cut are directly related.

Consider a thin, rectangular crystal-plate cut from a slab such as we have been discussing. When the cutting plane is normal to the Z axis, the resulting cut is called a Z cut, and a

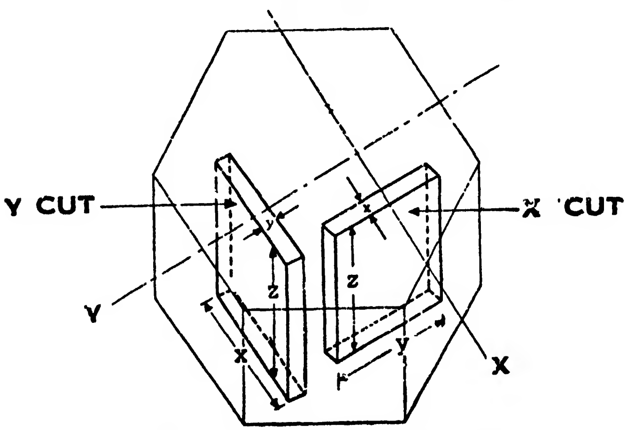


FIG. 266. X and Y Cuts in Quartz Slab.

plate so produced will have no piezo-electric properties if the electrodes also are normal to the Z axis, as has been explained.

When the cutting plane is made normal to an X, or electrical axis, the main faces lie in the YZ plane, and a parallelapiped so cut would have sides of lengths x, y and z, where these are dimensions parallel to the axes X, Y and Z, as shown in Fig. 266 (right). Such a crystal plate is known as an X cut; or sometimes as a "Curie," or "face perpendicular" cut, terms now to be deprecated. Piezo-electrically such a crystal will have a strong edge vibration along the Y axis, and a weaker mode of compression and extension along the X axis. Since the face dimensions are usually large compared with the thickness, the former frequency will be low compared with the

latter, the rule connecting frequency with crystal dimensions being approximately :

$$f_{\text{edge}} = \frac{2{,}750}{y} \quad . \quad . \quad . \quad . \quad (1)$$

$$f_{\text{thickness}} = \frac{2{,}750}{x} \quad . \quad . \quad . \quad . \quad (2)$$

where f is in kc/s and y and x in mm.

If we cut a similar thin rectangular plate lying in a plane normal to a Y axis, Fig. 266 (left), a Y-cut plate, sometimes known as a 30° plate, such a crystal will be found to be very active in the Y direction, that is in shear mode about Z, the resulting (thickness) frequency being :

$$f_t = \frac{2{,}070}{y}. \quad . \quad . \quad . \quad . \quad (3)$$

where f is in kc/s and y is in mm.

The reason for the 30° classification can be seen by reference to Fig. 262. From this diagram it is clear that rotation of a vertical cutting plane about Z alternates every 30° between an X and Y plane, and hence from an origin of an X cut, the adjacent Y cut will be obtained by rotation of the cutting plane by 30°.

The N.P.L. notation reclassifies X and Y-cut plates, as the vertical cutting plane is considered to have an origin along the X axis (giving a y cut) and thus a 30° rotation from this origin results in an X-cut crystal plate.

It should be noted in passing that certain workers consider adjacent X and Y axes as complementary, whereas others prefer to consider the complementary X and Y axes at right angles.

The formulæ 1, 2 and 3 are derived from a consideration of the mechanical structure of the crystal. The oscillation of a quartz plate is due to mechanical waves in the plate setting up stationary waves, either longitudinally or transversely to the mechanical axis.

The velocity of propagation for a mechanical wave is given by

$$v = \sqrt{\frac{E}{\rho}} \quad . \quad . \quad . \quad . \quad . \quad (4)$$

Where v is velocity in cms per second,

E is the elastic modulus in dynes per cm^2 in the plane considered.

ρ is density in gm per c.c.

For normal specimens of quartz

$$\left.\begin{aligned} E &= 8 \times 10^{11} \\ \text{and} \quad \rho &= 2{\cdot}654 \end{aligned}\right\} \text{Average values.}$$

Then $\quad v = 5{\cdot}5 \times 10^5$ cms per second.

Now $\lambda = \dfrac{v}{f} = \dfrac{5{\cdot}5 \times 10^5}{f}$ where λ is in cms and f is in cycles per second.

or $\quad \lambda = \dfrac{5{,}500}{f}$ where λ is in mms and f is kcs per second.

If the crystal is free, the fundamental vibration, whether longitudinal or transverse, will clearly be at a half wavelength with a node of movement at the crystal centre. Thus if y and x are crystal dimensions in mms along the Y and X axes, then since :

$$\lambda = 2y \text{ and } \lambda_1 = 2x$$

$$2y \text{ or } 2x = \frac{5{,}500}{f}$$

$$y \text{ or } x = \frac{2{,}750}{f} \quad \text{as stated above.}$$

This example applies to an X-cut crystal only.

Neither the X cut, nor the Y cut, crystal plates are now greatly used as both have a poor temperature-coefficient, and suffer from a phenomenon known as stepping, to be explained, due to the complex type of vibration within the crystal. With thin plates, where the thickness is not more than $\frac{1}{20}$th the edge length, the temperature-coefficient of the X-cut plate is of the order of -20 to -50 parts in 10^6 for each 1° C. temperature rise, and that of the Y cut, of the order of $+60$ to $+100$ parts in 10^6 for each 1° C. temperature rise, the thinner the plate cut the worse the value of temperature-coefficient.

Although the thickness frequency of a Y-cut plate varies approximately as $f_{kc/s} = 2{,}000/y$, yet if a plate be gradually decreased in thickness, a curve of f frequency against y does

not give a straight line but sudden jumps in frequency occur at points to give a discontinuous line. These discontinuities are termed stepping points, and result when the edge vibrations, controlled by the size of the faces of the plate, are coinciding with sub-multiple frequencies of the wanted thickness frequency.

We can really liken a quartz plate to a very high Q electrical resonant circuit (values of 25,000 in air and 200,000 in vacuo being possible), to which is coupled a number of subsidiary tuned circuits, also of high Q; the degree of coupling and frequency relationship to the fundamental varying with crystal dimensions. As long as these subsidiary circuits are in-harmonic, or have zero coupling, they will not affect the main oscillation, but small changes in dimensions, or in the method of holding the crystal, or a temperature change, may bring one or other of these circuits into effect.

As a plate is ground it is often possible to anticipate the step frequencies, and in the event of a step probably appearing at or near the required frequency, either the crystal must be utilised for a different frequency, or its outside dimensions varied so as to decouple or change the relative frequencies of edge to thickness. Avoidance of proximity to stepping points is also desirable on account of the fact that a crystal is usually less active in such a condition.

Crystals of Low Temperature-Coefficient

It is clearly an advantage if a crystal can be cut to have good piezo-electric properties and at the same time to have a low temperature-coefficient. There are many ways in which this object can be achieved, for a limited temperature range.

Since it is found that a Y-cut crystal has a positive temperature-coefficient and an X-cut crystal a negative one, crystals having a cylindrical or nearly cubic shape should have a coefficient which approximates to zero. This is found to be so. Thus Marrison produced low-temperature crystals by making them in the form of a thick ring the inner face of which was doubly tapered to a narrower neck at the ring centre.

The low temperature-coefficient is probably due to the general increase of thickness compared with the transverse

dimensions rather than by virtue of the ring construction, although the particular shape enabled Marrison to support the crystal by a line contact giving the least possible mechanical restraint.

T. D. Parkin has produced crystals which are nearly of cubic shape, cut to a dimension which compensates for the difference of temperature-coefficient.

The "Cube" is cut with sides exactly parallel to the *X*,

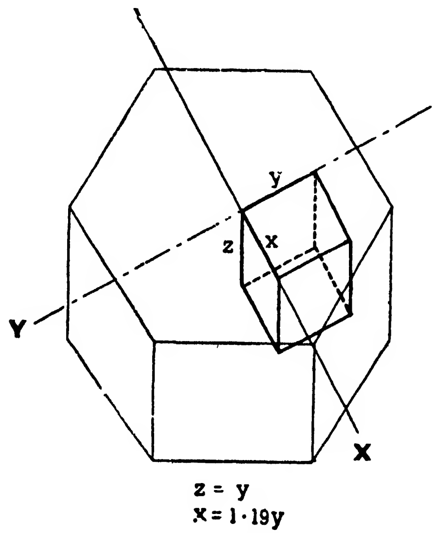


Fig. 267. Cubic Crystal.

Y and *Z* axes, the length along the *X* axis being approximately 1·19 times that along the *Y*, and the depth of the crystal along the *Z* axis being equal to that along the *X* axis (see Fig. 267). Not only does such a cut give a very low temperature-coefficient over a certain temperature range, but the dimensioning is such that the two modes of oscillation coincide and such a crystal is substantially free from stepping troubles.

Since the frequency of a Cubic crystal is given approximately by:

$$f_{\text{kc/s}} = \frac{1910}{y_{mm}}$$

it is seen that the range covered by reasonable size cubes

limits their use to the lower end of the high-frequency scale as shown in Table XX.

Inclined Angle Cuts

It has already been mentioned that X and Y-cut plates suffer from two disadvantages, a bad temperature-coefficient and spurious modes of oscillation. It has been found that plates can be cut at certain angles inclined to one or more of the axes which will be free from one, or other, or both these disadvantages, but it will be best to treat the two cases separately.

The resonant frequency of a crystal has been shown to be due to mechanical stationary waves set up within the quartz and from the formula on page 445 we can write

$$f_{\mathrm{kc/s}} = \frac{1}{2y}\sqrt{\frac{E}{\rho}} \quad . \tag{5}$$

where E is the elastic modulus in the plane being considered.

If a quartz plate is changed in temperature, the frequency

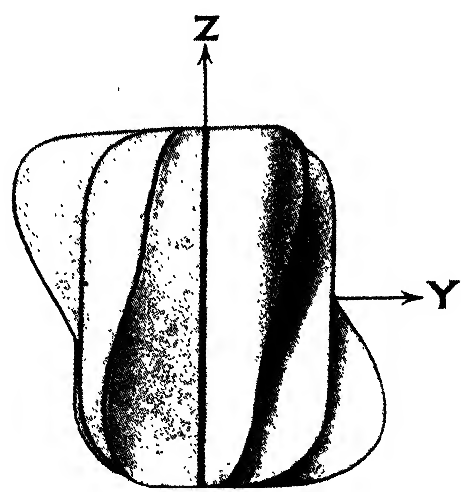


FIG. 268. Variation of Elastic Modulus in Quartz Crystal.

resulting will be dependent upon how the various terms in (5) vary, and it can be shown that whereas the temperature-coefficient of ρ is independent of orientation of the plate, that of the thickness dimension and of elastic constant vary with plate dimensions and orientation. The chief change brought about by change of orientation is as a result of the alteration in the elastic constant. Since quartz is non-isotropic, the elastic constant varies with direction, and in Fig. 268 is shown a three-dimensional diagram which indicates in polar form the value of E in the various directions, each alternate section considered as a polar curve depicting the change of E along

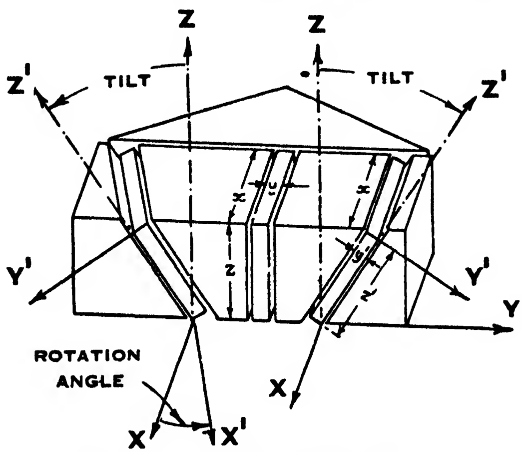


FIG. 269. Inclined and Tilted Plates.

the YZ and XZ planes. Thus the section lying in the plane of the paper is a YZ curve, the edge next adjacent forward is in an XZ plane, that forward again a YZ plane, and the section seen edgewise is in an XZ plane. From this it is seen that whereas the polar curve is symmetrical about an XZ axis plane (although not circular), it is very asymmetrical about a ZY axis plane, E reaching a value of about 13×10^{11} at 50° on one side, towards any m' face, and a minimum value of 7×10^{11} at 70° on the other. Viewing the crystal from different planes, this asymmetry reverses in direction at each adjacent Y axis. Thus Fig. 262 shows the polar curve in a YZ plane related to a left-hand and right-hand crystal respectively, but facing any positive X axis of the crystal, as seen by the plan and elevation

above. It should be clearly realised, however, that since a rotation through each 60° of the crystal before the observer reverses the direction of the sense of the axes, and changes the relative positions of *mm'* faces, it will also reverse the asymmetry of the elastic modulus figure. Thus the clockwise or anti-clockwise bias of these figures must not be associated with left- or right-handed crystals but only with a change of viewpoint and the relative positions of the *m* and *m'* face.

At and near the *Y* axis plane, that is in the region of greatest asymmetry, the modulus changes rapidly, and it is found that a large variety of cuts inclined at different angles with respect to the *XZ* plane have a substantially zero temperature-coefficient, over a limited temperature range.

Consider a rectangular section of quartz crystal cut from a

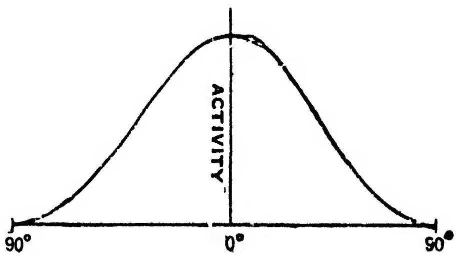


FIG. 270. Activity of Tilted Plate.

Z cut slab as shown in Fig. 269 (front section removed), and imagine a *Y* cut parallelepiped, whose dimensions are indicated by *x*, *y*, *z*, parallel to the *X*, *Y* and *Z* axes. As explained previously, such a slice will have an active shear vibration, and a less pronounced edge vibration. If we rotate the cut of such a plate, either way, about its bottom edge, i.e. about the *X* axis, Fig. 269 (right), the more it is rotated towards the horizontal position, the less piezo-electric activity it will have, the relationship of angle of rotation to activity being as shown in Fig. 270: owing to the asymmetric nature of the crystal structure, however, the temperature-coefficient tilt-angle curve for the shear mode assumes a form as shown in Fig. 271, full line. From this we observe that an angular rotation of 35° towards an *m* face or a reversed rotation of 49° towards an *m'* face results in a zero temperature-coefficient being obtained, both types of

vibration being still of the shear type, and both being liable to spurious oscillations, unless precautions are taken to avoid them. The 35° angle cut will have the greater activity, and it is desirable to point out that the sense of angles will change with the hand of the crystal and with the viewing position. By working with an *m* face as guide, however, the correct sense will always be obtained.

Crystals cut in this way, which were first investigated by

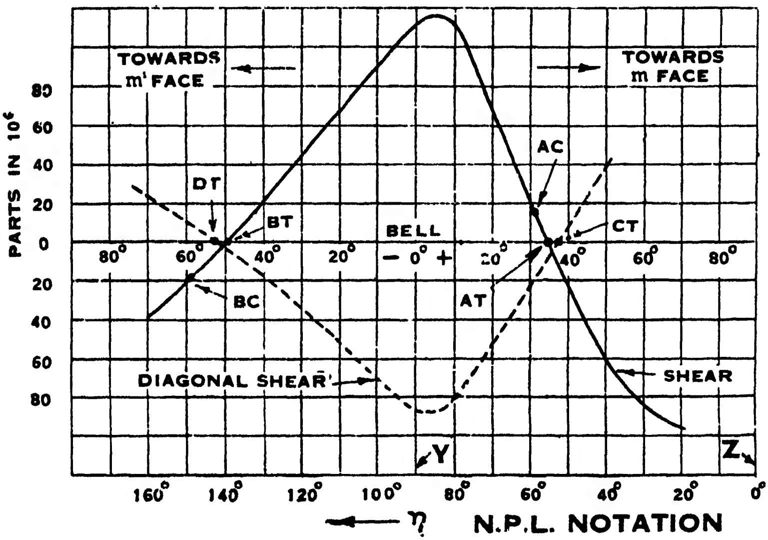


FIG. 271. Temperature-coefficients of Tilted Plates.

Bechmann and Koga, are known as *r* cuts by Koga, and *AT* and *BT* cuts by the Bell Telephone Laboratories.

With a tilted plate, since the crystal characteristics are changed, it is usual to denote dimensions x', y', z', and the new axes X', Y', Z', and thus with a plate tilted through 90° we get the anomalous conditions of a Y' axis coinciding with a Z axis, and a y' side coinciding with a z side, etc.

A simple inclined cut, such as we have described, can clearly be specified by two angles, one giving the inclination from a datum plane, either horizontal or vertical, and the other giving the angle of rotation about the Z axis from a given plane. Because of the asymmetry of crystal properties, it will

be necessary for sense to be specified, either in terms of the vertical or horizontal angle.

If, in addition to tilting the plate, we change its horizontal orientation by a bodily rotation about the Z axis (which will not change the angle of tilt), we get what is called by the Radio Corporation of America a VW cut, Fig. 269 (left), then for any given rotation angle W up to some 15° either side of the XZ plane it is possible to find a correct tilt angle V to give a zero temperature-coefficient. With all such inclined cuts it is assumed the final plate is cut from the inclined slice with sides parallel and normal to the top and bottom edges of the slice from which it was cut.

If the parallelepiped is cut from an inclined slice with its sides at an angle to the edges of the slice, as shown in Fig. 272 (bottom left), such a cut will be a cut " skewed " to all three axes. We could of course imagine such a plate as having been produced by giving to an original Y cut parallelepiped first an inclination about X, then a bodily rotation about Z, and then a tilt on to one corner, keeping its major face in the same plane.

The specification of a tilted skewed cut is rather more difficult. Its three axes can be specified in terms of three angles relative to arbitrary planes ; or the inclined slice from which the plate is finally cut can be specified in terms of two simple angles as stated, and the inclination defined separately.

N.P.L. Notation

The N.P.L. use, as a datum plane for the specification of all cuts, a plane at right angles to the Z axis, called the Z plane. Thus in Fig. 272 x, y, z is a Z cut parallelepiped cut at right angles to the Z axis (shown outside the slab). Any other cut is obtained by first giving to the cutting plane, now horizontal, a rotation about the X axis of η°, measured from $+ Y$ (that is, from the axis emerging from an m face), and then a bodily rotation of ζ° about the Z axis, the positive angle of rotation being from a positive X to a positive Y axis, the inclined cut being then simply defined as $\eta\ \zeta$.

With such a notation, the rotation angles are without the necessity of sense, positive or negative, because the vertical angle may have any value between 0° and 180°, and thus the asymmetry of the crystal is allowed for by whether the angle is

less or greater than 90°. Observe that the vertical angle is always measured from a positive Y axis, that is from one

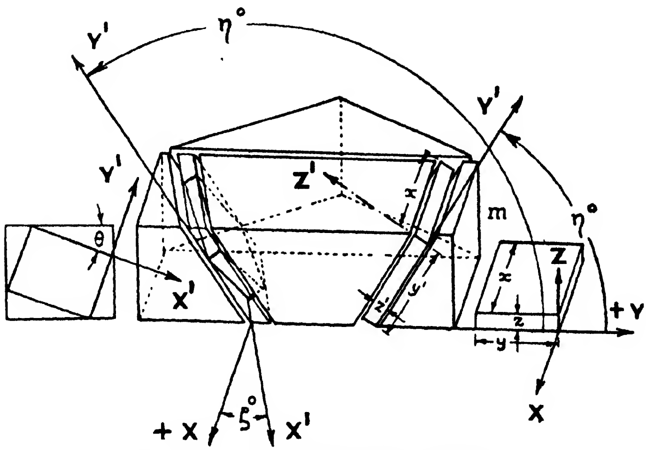


FIG. 272. Illustrating N.P.L. Notation.

emerging from an *m* face. This may be a clockwise or anti-clockwise direction depending not so much upon the hand of

TABLE XIX

Alphabetic	N.P.L.		Bell		R.C.A.	
			ANGLE OF			
	I.	*R.*	*I.*	*R.*	*R.*	*I.*
	η	ζ	ϕ	θ	*W. B.*	*V. A.*
Z	0°		90°		90°	0°
X	90°	30°	0°	60°	0°	0°
Y	90°	0°	0°	30°	0°	30°
AT	55°	0°	35°	90°	−35°	30°
BT	138°	0°	48°	30°	48°	30°
CT	39°	0°	39°	90°	−39°	30°
DT	145°	0°	55°	30°	55°	30°
AC	59°	0°	31°	90°	−31°	30°
BC	149°	0°	50°	30°	59°	30°
GT	51°	0°	45°			

NOTE.—In the above table *I* shows the angle of inclination and *R* the rotation angle.

the crystal as the viewing position, as previously mentioned. The horizontal angles may have any value up to 60° as rotation though this angle brings the cut back to a similar axis. Thus, using the N.P.L. notation, a *Y*-cut crystal becomes 90°, 0°, and an X-cut 90°, 30°, the relationship of the notation to this and other cuts being shown in Table XIX and Fig. 271.

In the case of a tilted skew cut, the angle of tilt of the final plate cut from the skewed slice is defined separately, namely as the acute angle (measured counter-clockwise) which one side of the plate makes with the top edge of the inclined slice, the inclined slice from which the plate is cut having been defined separately as above.

American Notation

In America two systems of notation have arisen, one originated by the Radio Corporation of America and the other by the Bell Telephone Laboratories. Both differ from the N.P.L. by adopting the *Z* axis as datum for the angle of inclination, and not the *Z* plane, and they differ from each other and from the N.P.L. by a different convention as regards the horizontal angle of rotation.

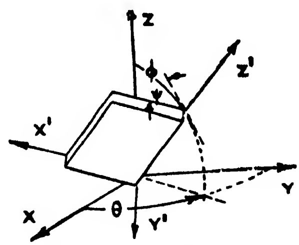


Fig. 273. Illustrating American Notation.

The Bell Telephone Laboratories identify their main types of crystal plates by letter notation and an inclination angle positive or negative, depending upon whether the new *Z'* axis is towards or away from an *m* face as indicated in Fig. 271. Whereas the inclination angles for the various alphabetic cuts are shown either side of the *Z* axis, their specification is, however, determined by an angular deviation from the three axis planes. Thus in Fig. 273, which shows a rectangular plate inclined to all three axes, its orientation is defined by the Bell Laboratories in terms of three angles ϕ, θ, ψ. Thus although in their alphabetic designation of *AT*, *BT*, etc., plates the new axis *Z'* is defined towards or away from an *m* face, in their specification the inclination angle is defined by $\phi°$ without sense. The asymmetry of the crystal is therefore allowed for

by giving sense to the rotation angle θ measured from an edge not a face, a positive sense being anti-clockwise, and the negative sense clockwise, for a right-handed crystal. As mentioned previously the Bell Telephone Laboratories definition of right-handed and left-handed quartz is opposite to that of other workers in the piezo-electric field.

The definition of the final tilt is obtained by the angle $\psi°$ which is defined as the angle by which a cut inclined $\phi°$ and rotated $\theta°$ about a horizontal axis, is then rotated about the new Z' axis. This second tilt will clearly change the direction of the X' and Y' axes from the direction before tilting, and it would appear therefore that this type of specification is somewhat difficult of interpretation.

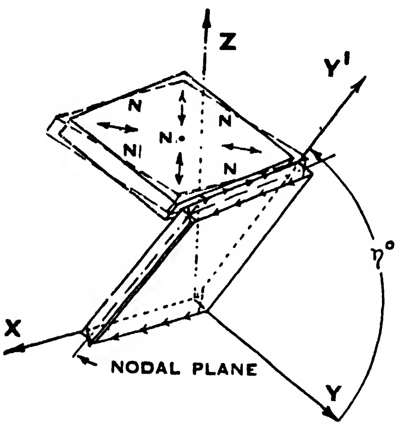


FIG. 274. Illustrating Vibration of A.T. and D.T. Plates.

The R.C.A. have a different system of notation. For the angle of inclination, they adopt the Z axis as datum and define the angle of inclination as $B°$, positive *away* from the *m* face, and negative towards it. For the rotation angle about Z, the specification is somewhat involved, but virtually it may be considered as specifying the angle (measured clockwise or anti-clockwise) from an equivalent Y axis. Thus a Y cut plate will be defined as A 30° B 0°.

Of these notations, correlated in Table XIX, the N.P.L. system is adequate for all types of cut, it is quite without ambiguity, and is the only system which defines a plate inclined to all three axes in a simple manner. Where no ambiguity exists, the designation of plates by letter appears highly desirable, such for instance as " X " plate, " Y " plate, etc.

The η 55° ζ 0° (AT) and the η 138° ζ 0° (BT) zero temperature cuts previously mentioned are thickness shear modes, suitable for high frequency work as indicated in Table XX. Since, however, a crystal has a twofold dimension at right

angles, it would be expected that two cuts complementary to 55° and 138° would also give a zero temperature-coefficient, but having a different mode of oscillation—namely, across the face of the plate. This is found to be so, and low-frequency, diagonal-shear-mode plates at values near η 145° (DT) and ζ 51° (LT) inclination are found to have a zero temperature-coefficient, the first being complementary to the AT plate and the second to the BT. The relationship of the AT and DT plates is shown in Fig. 274, which also indicates the modes of vibration, and the dotted curve in Fig. 275 shows the relation-

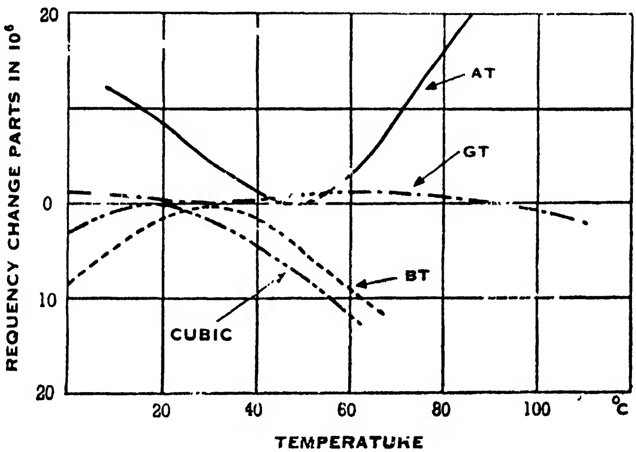


Fig. 275. Temperature-frequency Curves for Various Cuts.

ship of temperature-coefficient and angles of inclination of diagonal-shear plates.

All the plates we have mentioned suffer to some extent from coupling effects to other modes, and as mentioned previously the "temperature-frequency change" relationship does not hold over an infinite temperature range, but is limited. In most cases the shape of the "temperature-frequency change" curve is parabolic as shown in Fig. 275, where curves for selected crystal plates are given.

The Radio Corporation of America have found that with a change of rotation angle away from the X axis, and a corresponding alteration of the inclination angle, a whole series of

cuts, with appropriate angular notations (designated *VW*), can be found with an extended range of temperatures over which the coefficient is zero.

Zero-Coupling Crystal Cuts

It is found that a simple inclined cut, oscillating in H.F. shear mode, can be found where the coupling between edge and thickness modes is zero. Two such cuts are η 59°, ζ 0°, and η 149°, ζ 0°, so-called *AC* and *BC* by the Bell Telephone Laboratories, but by a change of rotation angle ζ, a series of cuts having similar characteristics will be found.

The Bell Telephone Laboratories have also devised a cut having a limited frequency range, which is most interesting as it not only has zero coupling, but a zero temperature-coefficient over a very wide temperature range. This cut is called by them *GT*, and in N.P.L. notation is a rotated inclined cut of η 51°, ζ 0°, θ 45°, and is obtained by first cutting a *CT* η 51°, ζ 0°, and then cutting from this a crystal plate with sides at 45° to the edges of the block. It has relative edge dimensions of width to length of ·86 to 1, and the width determines the frequency obtained.

According to the Bell Laboratories such a crystal oscillates in compression and extension about a nodal line across the centre of the plate. This can be seen from Fig. 274, where the top plate shows the diagonal shear vibration of a *CT* type plate, and it will be clear that if a plate was cut from this with sides at 45°, there will be extension and compression about a diagonal of the *CT* cut plate. The extended temperature range over which its frequency remains constant can be seen from Fig. 275.

The overall dimensions of the various plates we have mentioned are governed partly by power considerations and the use to which the crystal is to be put, partly by holder design and to avoid spurious oscillation modes, but for thin plates a face dimension of about 25 mm square appears to be a general dimension, the actual thickness being usually determined by the frequency required. The table following gives relevant information regarding the various types of cut we have enumerated.

TABLE XX

Cut.	Type of Oscillation.	Frequency kc/s given by :— (x, y, in mm.)	Frequency Range kc/s.	Frequency Change. Parts in 10^6 per 1° C.	Remarks.
X	Compression and Extension	$\frac{2{,}750}{y \text{ and/or } x}$	40— 1,000	−20 to −50	Active. Nearly obsolete for transmitters. Steps.
Y	Shear	$\frac{2{,}070}{y}$	500— 3,000	+66	Very active. Nearly obsolete for transmitters. Steps.
AT	Shear	$\frac{1{,}630}{y'}$	500— 2,000	Zero at correct temperature	Very active. In great use. Steps. Can be clamped.
BT	Shear	$\frac{2{,}500}{y'}$	2,000—15,000	„	Active. In great use. Steps. Can be clamped.
CT	Diagonal Shear	$\frac{3{,}100}{x}$	100— 200	„	Very active. Can be only clamped at centre.
DT	Diagonal Shear	$\frac{2{,}100}{x}$	70— 150	„	Very active. Can be clamped at centre.
GT	Compression and Extension	$\frac{3.292}{x'}$	60— 1,000	Zero over wide range.	Active. Clamp on centre.
AC	Shear	$\frac{1{,}620}{y'}$	500— 2,000	+20	Active. No coupling.
BC	Shear	$\frac{2{,}560}{y'}$	2,000—15,000	—20	Active. No coupling.
Cube	Shear	$\frac{1{,}925}{y}$	75— 750	Zero at correct temperature	Active. Free of steps. Clamp at critical points.
VW	Shear	$\frac{1.600}{y'}$ to	500— 2,000	„	Very active. Modification of AT and BT.
VW	Shear	$\frac{2.500}{y'}$	2,000—15,000	„	

Résumé of the Various Cuts

From a study of the above table it is clear that, by the selection of the proper cut, a great range of fundamental frequencies can be covered.

Not all these crystals are used, however, for oscillator work, as they are not active enough and would therefore be difficult to start into oscillation and probably give only very small output.

The most active plate is the *Y*-cut, but its bad temperature-coefficient precludes its use in most cases, and if we exclude the *X*-cut crystal for the same reason we can assume that of the others, the *AT*, *BT*, *CT*, *VW*, *DT*, *GT*, and Cubic are useful for oscillators. The *X*, *AC*, *BC*, and *GT* are used as resonators, and as impedances in filter networks.

In its range the Cubic crystal is the most economical as regards quartz and manufacturing costs. Not only is it the smallest in overall bulk, but because the faces are parallel to the principal axes of the crystal, it is easy to set up on the jig and cut almost straight away to finished dimensions. Further because of its shape a much greater percentage of the good quartz may be utilised. This type of crystal is free from spurious oscillations, but its temperature-frequency characteristic is parabolic. It is not critical as regards holder design and its small bulk makes for a compact unit.

Of the inclined cuts, the *AT*-cut is most active, it can be made for a large frequency range, and it gives a good output and can be clamped in its holder. It suffers, however, from stepping troubles and like all angle cuts it is wasteful of material in manufacture. The *BT*-cut can be made to higher frequencies than any other type of crystal, it can be clamped, but it is less active than the *AT*-cut and gives smaller output. The *CT* and *DT* cuts cover a lower frequency range and are alternative to the Cubic type of crystal. They require careful clamping at the centre, thus requiring a special type of holder, and they cannot be overloaded. One of the great advantages is, however, that they can be ground to a definite frequency at a definite temperature without great difficulty. This is possible because the frequency of the plate can not only be decreased by grinding, but also increased as well. Increase of frequency is obtained by grinding the corners away, and decrease of frequency is possible by grinding away the centre,

and thus a correction for over-grinding can be made without difficulty, although it will be realised that the temperature-frequency curve is parabolic.

The *GT* plate appears to be suitable for all classes of work and its mode of oscillation lends itself to rigid clamping by a holder of simple design.

Crystal Holders

It will be obvious that the crystal holder will play some part in determining the final frequency of the crystal unit as a whole. Although the crystal itself is the main factor for determining the frequency of oscillation, the method of mounting must influence the frequency, even though this will be a second order effect, as the holder is bound to add capacitance and damping. Thus, although a simple and inexpensive type of holder can be devised for crystals having a wide tolerance, such, for instance, as used in ship stations, portable sets, and rough-check wave-meters, where a precision of not more than 1 part in 20,000 may be required, the design of a holder for crystals used for high precision work (such, for instance, as in a frequency-checking station) will need the most careful consideration.

Before modern methods of grinding and surfacing to a precise value were developed, it was common practice to carry the crystal in a holder containing an air gap, which could be adjusted to bring the crystal to the required frequency, a variation of 1 part in 2,000 being easily possible by gap variation with the ordinary *Y*-cut crystal. This is not good practice where only a small tolerance is possible, and present-day holders are usually designed either without gaps, or when a gap is used it will be fixed or possibly given the very smallest amount of adjustment. For the highest precision, the holder will often be evacuated, or partially evacuated, as this protects the crystal from moisture, barometric changes have no influence, and the vibration of the crystal cannot set up air resonance changes, but generally speaking such crystals will not be used for oscillator work.

Holders generally are of two main types. One which clamps the crystal either on the faces of the crystal, where this is possible, or nodally, on point, or line contacts depending upon

the type of crystal being held. Secondly, a type of holder having a fixed gap (or possibly having a very small variation) and designed to carry the crystal resting on one face and lightly constrained laterally. With the Cubic crystal another type of holder is sometimes used in which the crystal is suspended in air between the electrode faces.

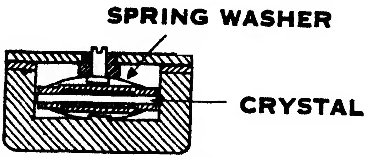


FIG. 276. Clamped Crystal Holder.

A simple type of holder, in which the crystal is clamped lightly between two electrodes by means of contact springs, is shown in Fig. 276, such a holder being useful for oscillator crystals, using an *AT* or *Y* cut.

Figs. 277 and 278 show two types of holders designed for Cubic cut crystals. In the first a small air gap is allowed between the top of the crystal and the electrode, and the crystal is held laterally inside a thin ring, notched to take the corners of the crystal. Such a crystal unit would be suitable for tolerances up to 2 parts in 10^6, and would be used for oscillators. The holder shown in Fig. 278 is novel as the crystal is suspended in a cradle of silk threads, with threads stretched across the faces of the crystal to prevent sideways movement, the crystal and cradle assembly being mounted centrally between the electrode faces so as to leave a small air gap each side. Such a crystal assembly has a precision within 3 parts in 10^7 and would be of use for frequency-checking equipment.

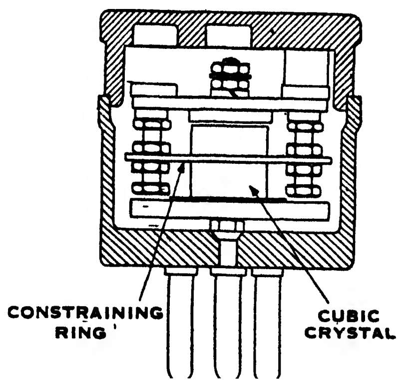


FIG. 277. Holder for Cubic Crystal.

For precise work with certain of the inclined cuts, clamping can be resorted to, but only at selected nodal points. It is clearly a great advantage if the design can be such that firm

clamping is possible, as it solves the transport problem very considerably, for with unclamped crystals any mechanical disturbance tends to shift the frequency. For instance, in the *GT* cut, where the crystal vibrates about a nodal line across the centre, a holder is used which grips the crystal between two wedge-shaped jaws along the nodal line. This is clearly a very simple type of holder to design. Another method of clamping is between hardened pointed steel pins, nodal points being chosen, which may be on the face as in the case of the *CT* and *DT* cuts, or along the edges as in the case of *BT* cuts.

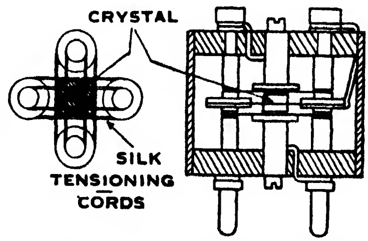


FIG. 278. Holder for Cubic Crystal.

The great variety of cuts which appear to be possible having either a zero temperature-coefficient, or small coupling between modes, or both, is so large, that the technique of crystal production and holder design is clearly only now in an early stage of development, and we should see considerable advances in the next few years.

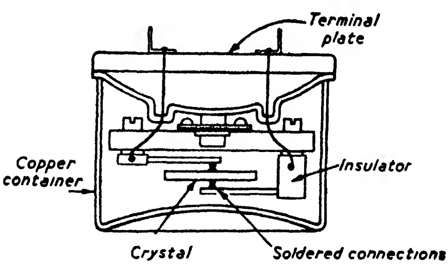


FIG. 279. Holder for *GT* Cut Crystal.

The British Post Office has developed, for frequency standards, the use of deposited gold electrodes. In some cases contact to these is made by point contacts under pressure. In the most recent types, however, the connections are soldered on to small silver spots on the centres of the gold electrodes.

In Fig. 279 is shown the holder for a 100 kc/s standard, using a *GT*-cut crystal. The copper can is evacuated to avoid changes of frequency brought about by changes in barometric pressure and to make air damping very small. The can is contained in a temperature-controlled oven.

Circuit of the Valve-Maintained Quartz Crystal

Any mechanical oscillator can be analysed in terms of an equivalent electrical circuit, which would behave in the same way. An equivalent circuit for a quartz slice in its holder is shown in Fig. 280 and it will be seen that the slice itself is equivalent to a series circuit of very high inductance and small capacitance, whilst C_1 represents the capacitance between the electrodes, with the crystal in place, but not oscillating.

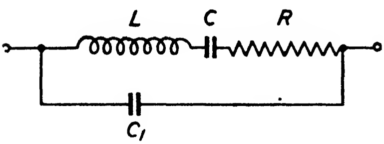


FIG. 280. Equivalent Circuit of Quartz Slice.

As an example of circuit values, one of the crystals mounted by the Post Office in the manner just described is represented by $L = 22{\cdot}4H$, $R = 24\Omega$, $C = 0{\cdot}113\mu\mu F$, $C_1 = 50\mu\mu F$, $Q = 585{,}000$. Such a crystal has been specially selected to act as a standard and is oscillating in a vacuum. The Q of an average crystal, as used to drive a transmitter, would be much lower than this—perhaps two or three thousand—but this is still much higher than could be realised in a practicable LC circuit.

Such a circuit will show a very low impedance (pure resistance) at a frequency which makes L, R and C come into series resonance. The impedance will then rise very rapidly to a very high value when the whole circuit (including C_1) comes into parallel resonance.

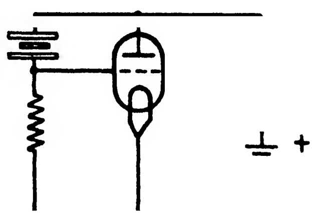


FIG. 281. Pierce Crystal Oscillator.

The circuit of Fig. 281 is in general use for maintaining the crystal. Considering this circuit in conjunction with the equivalent circuit of the crystal, it will be seen that we have a common-anode type of oscillator and the frequency will be almost exactly that for which the crystal circuit comes into parallel resonance. The LC circuit between anode and cathode will only have a very small control on the oscillation frequency and its resonant frequency must be below that of the crystal, in order that it may form a capacity reactance.

If the circuit of Fig. 282 is considered, together with the

equivalent circuit, it will be seen that it will function in a similar way to the tuned-anode, tuned-grid oscillator, the frequency being almost exactly the parallel-resonant frequency of the crystal. Here, again, the *LC* circuit will have only a very small effect on the frequency, but in this case it must be tuned

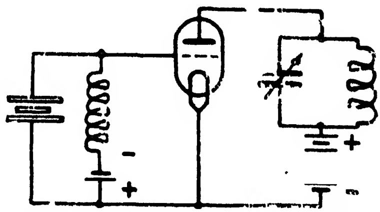
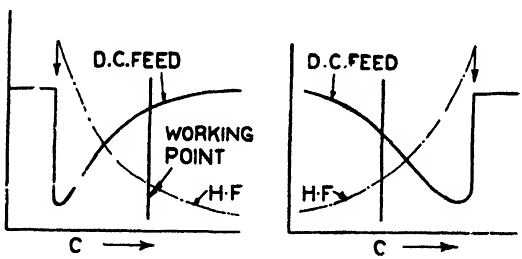


FIG. 282. Miller Crystal Oscillator.

to a higher frequency than that of the crystal in order that the anode circuit may be inductive.

With both types of circuit it is undesirable that the tuning of the *LC* circuit should be too near that of the crystal, as its



a *b*

FIG. 283. Illustrating Effect of Tuning the Anode Circuit.

effect upon the oscillation frequency will then be more marked. Fig. 283 shows how output and anode feed varies for both circuits, as the tuning of the anode circuit is varied.

Bridge-Stabilised Oscillator [12]

A crystal circuit for high-precision work, suggested by Meacham, is the bridge-stabilised oscillator, in which the amplitude of oscillation, as well as the frequency, is automatically kept constant by a special bridge.

The circuit is shown in Fig. 284, where R_1 is a resistance the value of which depends upon its temperature (usually a tung-

CHAPTER XII

ELECTRON OSCILLATORS

THE consequences of the fact that electrons take a finite time to travel between cathode and anode of a valve have already been discussed in Chapter X.

Although, by modifying valves and circuits, oscillators employing feed-back have been made to work at increasingly high frequencies, the highest frequencies in use (such as 10,000 Mc/s) are generated by utilising transit-time effects, instead of reducing them as much as possible.

Oscillators of this type may be termed electron oscillators. We have included in this chapter, as a matter of convenience, the dynatron oscillation of the magnetron, although this is not dependent upon transit time.

Many investigators have worked on electron oscillators and the literature is very extensive. Due to the high frequencies generated, measurements are limited and approximate and the interpretation of the complicated experimental data correspondingly difficult. Remarkable progress in measuring technique has been made in recent years, however. Calculations of electron paths in practical valves are complex and tedious. As a consequence, conflicting theories have been put forward from time to time.

The need in radar for powerful transmitters and sensitive receivers working at frequencies of thousands of megacycles per second stimulated experimental and theoretical work on electron oscillators and very great progress was made in their development and practical design.

When discussing oscillators in the previous chapter we considered them as amplifiers with feed-back. Alternatively, we can regard the valve as a means for producing a negative resistance in the resonant circuit, and so cancelling out the inherent positive resistance. Thus any shock to the resonant circuit produces an oscillation which does not die down, as it

would without the valve circuit, but builds up to a value dependent upon the valve characteristic.

This view of an oscillator is in some ways more comprehensive than the amplifier-with-feed-back idea, because some oscillators (the arc or tetrode, for example) can be made to produce a negative resistance by virtue of their volt/ampere characteristics, without any feed-back circuit.

Any generator can be regarded as comprising a negative resistance because the current has a component in phase with the voltage, instead of in phase opposition, as in an ordinary resistance. Hence power is being given out instead of dissipated.

When we are dealing with frequencies so high as to be comparable with transit time, then the possibility arises of producing a negative resistance at these frequencies, even though the static characteristic of the valve does not exhibit it.

If we can arrange to drive a cloud of electrons in opposition to the oscillating electric field between the electrodes, then they will be retarded and will yield up energy which can sustain the oscillations. There will inevitably be electrons which are accelerated by the oscillating field and therefore abstract energy. Sustained oscillations will only be possible, therefore, if we can make these electrons abstract less energy than the others give up, and good efficiency will only be possible if we can make them abstract much less energy.

All electron oscillators depend, therefore, on favourable interaction between electrons in motion and electric fields, though it is not always easy to understand the exact mechanism by which oscillations are maintained.

Barkhausen-Kurz Oscillator

In 1919, Barkhausen and Kurz, whilst testing the vacuum of transmitting valves, placed a positive voltage on the grid and a small negative voltage on the anode. They found that a current was recorded in the anode circuit and, by attaching a Lecher-wire circuit between grid and anode, ascertained that a short-wave oscillation was taking place. The frequency was mainly controlled by the grid voltage and the oscillations were not dependent upon reaction in the external circuits.

We can best explain the Barkhausen-Kurz oscillation by

considering a triode valve having a plane parallel-electrode system preferably such that the distance between cathode and grid is approximately the same as between grid and anode, the grid being of rather open mesh.

Assume a positive potential $+ E$ applied to the grid and let the anode be at zero potential. This means we have similar potential gradients rising from both cathode and anode to the central grid. Since the cathode is the emitter, we shall obtain a D.C. grid current I_g, and the heat appearing at the grid is due to the energy given up by the kinetic energy of the electron stream, and is a measure of the work done on the electrons by the accelerating field between cathode and grid. Generally speaking many of the electrons will stream direct from cathode to grid, but in certain valves it is found that quite a large percentage of them will be found to execute an oscillatory movement within the valve before capture. This is because of the probability of an electron missing the grid in its passage; the consequent retarding field in the grid-anode space will cause the electron to come to rest at a point just short of the anode, from whence it will be accelerated back towards the grid once more. We can imagine certain electrons therefore executing a "to and fro" shuttle movement before final capture. The maximum velocity such electrons achieve and the frequency of oscillation will depend on the value of E and the distance from cathode to anode primarily, whereas the average number of oscillations made will probably depend mostly upon the grid mesh.

We will calculate the time taken for an electron to make a complete trip from cathode to anode and back, carrying out the calculation in the M.K.S. units.

Q = Charge on an electron in coulombs ($1 \cdot 59 \times 10^{-19}$).
m = Mass of an electron in kg ($9 \cdot 04 \times 10^{-31}$).
E = Grid potential in volts.
l_g = Distance from cathode to grid.
l_a = Distance from cathode to anode.
v = Velocity of electron (metres per sec.) when it shoots through the grid.

If it is assumed that the electron leaves the cathode at zero velocity, then the work which has been done on it when it

reaches the grid is QE (kg–m) and this equals its kinetic energy, so that

$$QE = \tfrac{1}{2} mv^2 \quad . \quad . \quad . \quad . \quad (1)$$

and the maximum velocity $v = 5{\cdot}93 \times 10^5\sqrt{E}$ metres per sec.

As the field in such an extended, parallel-plane arrangement will be uniform, the average velocity

$$\frac{v}{2} = 2{\cdot}97 \times 10^5\sqrt{E} \text{ m/s} \quad . \quad . \quad . \quad . \quad (2)$$

Hence the time taken to travel from cathode to grid is

$$\frac{l_g}{2{\cdot}97 \times 10^5\sqrt{E}} \text{ secs.}$$

and from grid to anode is

$$\frac{l_a - l_g}{2{\cdot}97 \times 10^5\sqrt{E}} \text{ secs.}$$

Thus the time for a complete oscillation from cathode to anode and back will be

$$\frac{2l_a}{2{\cdot}97 \times 10^5\sqrt{E}}$$

and the frequency, f_o, of the oscillation

$$= 0{\cdot}149 \frac{\sqrt{E}}{l_a} \text{ Mc/s} \quad . \quad . \quad . \quad . \quad (3)$$

The corresponding wavelength will be

$$\lambda = \frac{{\cdot}3 \times 10^8 \, l_a}{0{\cdot}149 \times 10^6\sqrt{E}},$$

$$= 2010 \frac{l_a}{\sqrt{E}} \text{ m.} \quad . \quad . \quad . \quad . \quad (4)$$

Note that both λ and l_a are in metres.

The valves usually used for electron oscillators have a single straight filament down the centre of a cylindrical, co-axial grid and anode. It is evidently mainly a geometrical problem to extend the simple theory just given, to meet this case, but the above formulæ are sufficient to enable an estimate of the oscillation frequency to be made.

The above relationships have been deduced for a single

electron, but it is evident that when detectable oscillations are being produced there must be a vast number of electrons carrying out this oscillating motion; in other words, there must be an oscillating space charge between the electrodes. This distorts the potential distribution so that the equivalent position of cathode and anode are closer to the grid and in consequence the wavelength found experimentally is usually less than that given by the above expression.

Since the electrons are leaving the cathode in random fashion it would appear that the net effect of the oscillations would be zero, but it is found possible to co-ordinate their motion so as to extract a small proportion of the oscillating energy they possess and which they have derived from the D.C. supply.

Suppose that we assume a small alternating E.M.F. applied between grid and anode, the period of this E.M.F. being equal to the period of oscillation of the electrons. Then the electrons which are accelerated as they shoot through the grid will arrive at the anode and there will be anode current. These electrons have absorbed energy from the alternating E.M.F. but it will be seen that they have been quickly removed from the circuit.

Electrons which are retarded as they pass through the grid will fail to reach the anode, and will turn back again towards the grid. They will be still retarded, however, because the E.M.F. will have reversed in sign. Hence these electrons will execute smaller and smaller oscillations until they are captured by the grid. These electrons have evidently been giving up energy to the applied E.M.F. and it will be seen that they remain longer in the cathode/anode space. Thus it is possible for energy to be fed into the source of the alternating E.M.F. rather than extracted from it. If a resonant circuit is connected between grid and anode, therefore, oscillations will be maintained in it if the energy supplied from the electron motion exceeds the losses in this circuit.

We worked out the time taken for an electron to make a complete journey from cathode to anode and back, but, if the distance travelled by the electrons varies as the oscillations decrease in amplitude, then the transit time is different and the electron oscillations tend to get into the wrong phase with respect to the circuit oscillations. This would appear to be

one of the reasons why the efficiency of positive-grid oscillators is invariably low.

In order to obtain good results from a B–K oscillator it is necessary to employ a circuit of very high Q, such as a resonant line. Such a circuit is shown in Fig. 285, the wavelength being adjusted by the position of the slider.

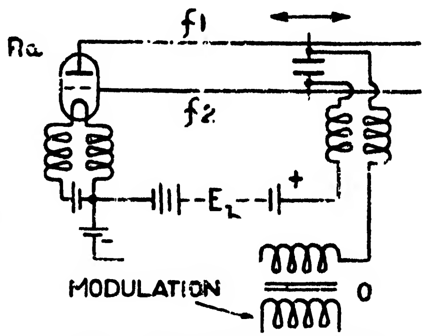


FIG. 285. Circuit for B.K. Oscillator.

Valves used for this purpose are usually made with a tungsten, bright-emitting filament, as this is the most satisfactory type to stand up to electron bombardment. Even with a tungsten filament this tends to shorten the valve life and there is also trouble with the grid overheating. The cathode temperature has to be carefully adjusted to maintain oscillations and this may have something to do with the control of the number of oscillations the useful electrons make before capture.

Some special types of positive-grid oscillator have been developed but the performance of the electron oscillators to be described later in this chapter is so superior to that of the positive-grid oscillator that it seems unlikely that they will be used to any extent in the future.

The Magnetron

This is a diode valve having, in its simplest form, a cylindrical anode held at positive potential, coaxial with a straight filament. The essential feature is that the valve is placed in a uniform magnetic field directed along the axis of the anode so that the electrons emitted from the cathode are influenced by both electric and magnetic fields.

An electron moving in a magnetic field experiences a force given by $BQv \sin \phi$ (where B is the magnetic flux density, Q is the charge on an electron, v is the velocity and ϕ is the angle between the directions of B and v). The force is perpendicular to the direction of v.

In consequence, electrons leaving the cathode will follow curved paths, such as those shown in Fig. 286, and it will be

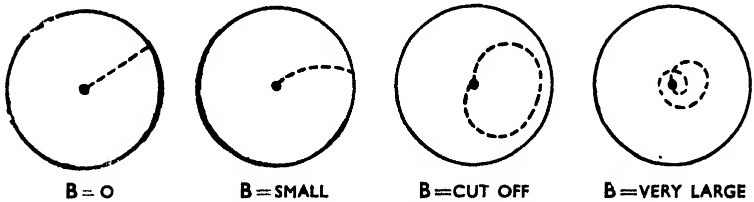


FIG. 286. Electron Paths in Full Anode Magnetron.

seen that when B is increased above a certain critical value, termed " cut-off," the anode current ceases and the characteristic will be as in Fig. 287.

The magnetron was first produced by Hull in 1920 and he proposed to use it as a kind of relay. If the magnetic field was provided by a current through a solenoid and this current

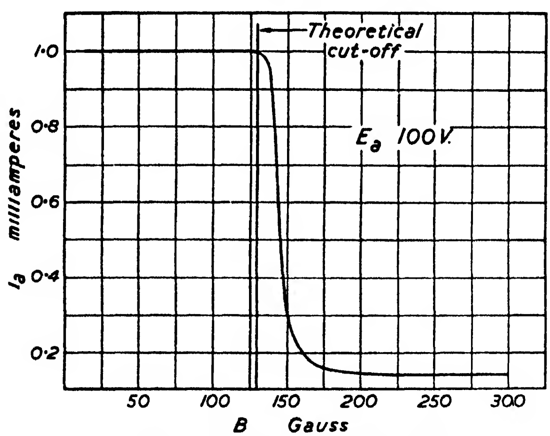


FIG. 287. I_a/B Curve for Magnetron.

was adjusted to very near cut-off, then a small additional current from a control circuit would cut off the anode current. Control of anode current in this way proved much less convenient than by grids and this use of the magnetron never became important.

In 1924, however, Zăcèk showed that very high-frequency oscillations could be produced in a magnetron in which the

magnetic field was maintained constant and the use of the magnetron in this way was studied by many workers. Its use in actual wireless equipment was rare, however, until in 1940 Randall and Boot and others developed a special type which produced efficiently the short pulses of very high frequency and very high peak-power, required for centimetric radar.

In most magnetrons the anode is divided into two or more segments and, in this case, three types of oscillations are possible. The names given to these by different workers are rather varied but " electronic," " resonance " and " dynatron " seem to be the more common terms.

Electron Motion in a Parallel-Plane Magnetron

Although a magnetron having parallel, plane electrodes, such as that indicated by Fig. 288 would not be practical, it is

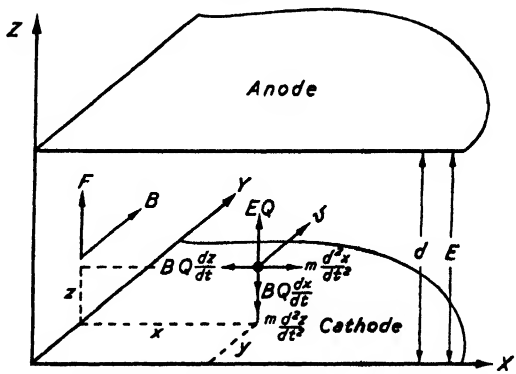


FIG. 288. Parallel-Plate Magnetron.

instructive to work out the path taken by an electron in such an arrangement, because the solution is much simpler than that for the cylindrical case and the same kind of effects occur.

The electric field will be uniform and will be along the Z axis, whilst the magnetic field is along the Y axis. An electron leaving the cathode at zero velocity will be accelerated in the Z direction by a force FQ but will have a force BQv upon it, normal to its direction of motion. Because the velocity of the electron is changing, there will be inertia forces given by the product of mass and acceleration.

These forces are shown in Fig. 288, resolved into components along the axes. Since the force due to the magnetic field is normal to the direction of motion, it will be seen that if the component of velocity along the X axis is $\frac{dx}{dt}$ this will produce a force along the Z axis of $BQ\frac{dx}{dt}$.

There being no force along the Y axis the motion will be entirely in the XZ plane if the electron is assumed to leave the cathode normally.

The equations of motion are seen to be

$$m\frac{d^2z}{dt^2} = FQ - BQ\frac{dx}{dt} \qquad . \quad . \quad . \quad (5)$$

$$m\frac{d^2x}{dt^2} = BQ\frac{dz}{dt} \qquad . \quad . \quad . \quad . \quad . \quad (6)$$

Since electric and magnetic quantities are involved, we have to consider carefully the system of units to employ. If we use the M.K.S. system, then the above equations are consistent without any conversion factors.

The solution to these equations can be written

$$x = \frac{F}{B\omega}(\omega t - \sin\omega t) \qquad . \quad . \quad . \quad (7)$$

$$z = \frac{F}{B\omega}(1 - \cos\omega t)\,. \qquad . \quad . \quad . \quad (8)$$

where $\omega = \frac{BQ}{m}$.

(This can be checked by differentiating the equations and combining them.)

These equations are those of a cycloid generated by a point on the circumference of a circle of radius $F/B\omega$ rolling on the cathode plane with angular velocity ω. The time for one revolution of the rolling circle is $2\pi/\omega$ and this is seen to be the time for one flight of an electron from cathode and back.

The mean velocity in the x direction is given by

$$F/B \text{ metres/sec.} \qquad . \quad . \quad . \quad . \quad (9)$$

The maximum travel in the Z direction is evidently given by the diameter of the rolling circle, that is by

$$\frac{2F}{B\omega} \text{ or } \frac{2Fm}{B^2Q}.$$

If we make this equal to d then we get the cut-off conditions,

$$B_c^2 = \frac{2Fm}{Qd}$$

For this uniform electric field, $F = E/d$ and hence

$$B_c = \frac{1}{d}\sqrt{\frac{2Em}{Q}} \quad . \quad . \quad . \quad . \quad . \quad (10)$$

If E is in volts, Q in coulumbs, m in kilogrammes, d in metres and B in webers per square metre, as required by the M.K.S. system, then

$$B_c = 3{\cdot}37 \times 10^{-6} \frac{\sqrt{E}}{d} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (11)$$

$$f_c = \frac{B_c}{2\pi} \cdot \frac{Q}{m} = 2{\cdot}81 \times 10^4 B_c = 9{\cdot}45 \times 10^{-2} \frac{\sqrt{E}}{d} \text{ Mc/s} \quad . \quad (12)$$

where $\frac{1}{f_c}$ is the time taken by an electron to journey from cathode to anode and back.

The wavelength of any radiation produced would be given by

$$\lambda = \frac{1{\cdot}07 \times 10^{-2}}{B_c} = 3180 \frac{d}{\sqrt{E}} \text{ metres} \quad . \quad . \quad (13)$$

If we wish to use the gauss (lines per square cm) as the unit of B we shall need to multiply B by 10^4 and if we measure d in cm, then the equations become

$$B_c = 3{\cdot}37 \frac{\sqrt{E}}{d} \quad . \quad . \quad . \quad . \quad . \quad (14)$$

$$f_c = 2{\cdot}81\, B_c = 9{\cdot}45 \frac{\sqrt{E}}{d} \text{ Mc/s} \quad . \quad (15)$$

$$\lambda = \frac{10{,}700}{B_c} = 3180 \frac{d}{\sqrt{E}} \text{ cm} \quad . \quad (16)$$

For example, if $d = 0{\cdot}2$ cm, $E = 2{,}500$ V, then B would be 843 lines/cm to give the cut-off condition. The frequency produced would then be 2,360 Mc/s and the corresponding wavelength 12·7 cm.

Electron Motion in a Cylindrical Magnetron

The equation of motion of a single electron leaving the cathode of a cylindrical magnetron can be accurately calculated by similar methods to those given for the parallel-plane magnetron but the fact that the electric field is no longer uniform complicates the analysis considerably. Typical paths have already been given in Fig. 286.

The cut-off relation can be shown to be

$$B_c = 13{\cdot}4\,\frac{\sqrt{E}}{d} \quad . \quad . \quad . \quad . \quad (17)$$

where E = anode/cathode potential (volts)
B = field strength in lines per cm (gauss)
d = anode diameter (cm).

The presence of a large number of electrons will produce a space charge and therefore modify the electric field distribution but it can be shown that this does not modify the above relationship.

Electronic Oscillations in a Magnetron

It is found that if the magnetron is adjusted to cut-off and a circuit having a resonant frequency $\omega_c/2\pi$ is placed between anode and cathode, then oscillations are maintained in this circuit.

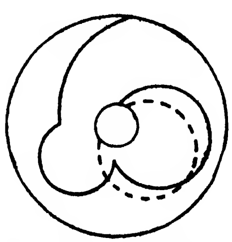


FIG. 289. Typical Electron Path during Electronic Oscillations.

Let us assume that a small p.d., $e_1 = E_1 \sin \omega_c t$, is superimposed on the D.C. between anode and cathode. Consider an electron which leaves the cathode when e_1 is just becoming negative, so that e_1 opposes the movement of the electron towards the anode. It fails to reach the anode and as it turns back e_1 also reverses in sign, so that it is still retarding the electron. Hence the path of the electron is a series of diminishing "scallops" as shown in Fig. 289. Evidently this electron has been delivering up energy to the source of e_1 during its whole flight.

If we now consider an electron leaving the cathode when e_1 is just becoming positive, this will be accelerated by e_1 and

will reach the anode. This electron has been extracting energy from the source but has been quickly removed.

We have considered the most favourable and the least favourable instants for an electron to leave the cathode. The majority of electrons will be between these two extremes but it will be seen that the electrons which are drawing energy from the source tend to be removed quickly and, therefore, the net result is to deliver energy.

Hence it is not necessary to supply e_1 from an external source but oscillations can be sustained in a resonant circuit.

If the retarded electrons are not removed from the electrode space after their oscillations have died down, they will be set into oscillation again by the R.F. field and will, therefore, extract energy from the circuit.

In the cylindrical magnetron, the axis of the magnetic field is "tilted" through a small angle (usually about 6°) with respect to the axis of the valve. This results in the electrons following spiral paths and emerging from the electrode structure. The tilt is adjusted experimentally to the value which gives the maximum output and, apparently, what is being done is to remove the optimum number of electrons when their useful "life" is over.

As an alternative to tilting the magnetic field, an end-plate, or disc, may be placed at each end of the electrode structure and a small D.C. potential applied to them.

It is found that the efficiency of electronic oscillations is not high and it would appear that many of the electrons absorb and deliver about equal amounts of energy to the circuit.

The time of transit of a single electron from cathode to anode and back can be calculated for cut-off conditions and the frequency of oscillation is found to be

$$f = 2{\cdot}44\,B \text{ Mc/s} \quad . \quad . \quad . \quad . \quad (18)$$

or the wavelength,

$$\lambda = 12{,}300/\mathrm{B} \text{ cm} \quad . \quad . \quad . \quad . \quad (19)$$

(where B is in gauss)

By using the relationship between E and B at cut-off, this can be written

$$f = 3{\cdot}26\,\frac{\sqrt{E}}{d} \text{ Mc/s} \quad . \quad . \quad . \quad (20)$$

$$\lambda = 920 \frac{d}{\sqrt{E}} \text{ cm} . \qquad . \qquad . \qquad . \qquad (21)$$

When a great number of electrons are oscillating to and fro, we shall evidently have an oscillating space charge and this will modify the transit time. A complete solution is rendered difficult but it is found experimentally that

$$f \simeq 2{\cdot}73B \text{ Mc/s} \qquad . \qquad . \qquad . \qquad . \qquad (22)$$

$$\lambda \simeq 11{,}000/\mathrm{B} \text{ cm} \qquad . \qquad . \qquad . \qquad . \qquad (23)$$

The electronic type of oscillation can also be produced in a magnetron having its anode split into two halves and this is the arrangement usually used, the oscillating circuit being connected between the two segments. The R.F. field will now be tangential, rather than radial, but the maintenance of

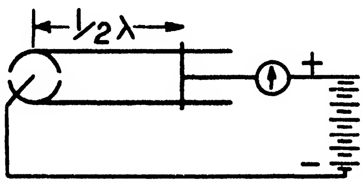


FIG. 290. Oscillator Circuit for a Split-anode Magnetron.

oscillations is still due to the interaction between this field and the circular component of the electron's motion. The frequency is, therefore, the same as for the full anode.

It is evident that the wavelength of these oscillations will be of the centimetre order and hence a line type of output circuit is most suitable, as shown in Fig. 290. The wavelength is, of course, mainly determined by the valve adjustments discussed above but varies somewhat with circuit tuning. The optimum output is obtained when the circuit tuning is adapted to suit the particular valve and valve adjustments. Fig. 291 shows a typical result with a magnetron having an anode of 1 cm diameter. It will be observed that, as the wavelength is reduced, it becomes necessary to increase anode voltage and magnetic field (as the equations show) and also filament current. It will be found that these adjustments are very critical, especially filament current.

Although the efficiency of a magnetron producing electronic oscillations is only about the same as that of a positive-grid

oscillator, greater outputs appear to be possible and shorter wavelengths can be obtained. This is because the absence of a

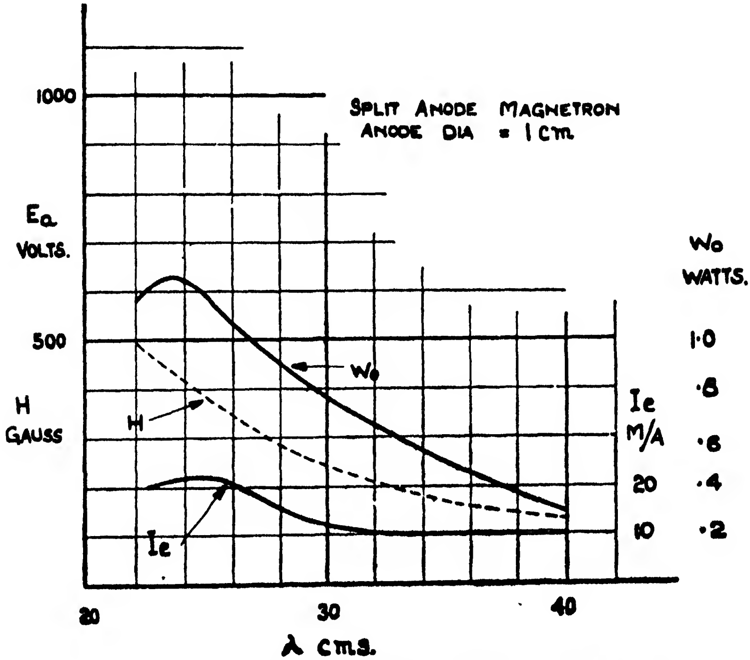


FIG. 291. Experimental Performance of Split-anode Magnetron.

grid makes a very compact electrode system possible and there are no difficulties due to the grid overheating.

A frequency of 61,000 Mc/s (0·49 cm) has been obtained by Richter, though the output was only about $2{\cdot}5 \times 10^{-7}$ watts for an input of 2·4 watts.

Resonance Oscillations in a Magnetron

This type of oscillation, also termed a travelling-wave oscillation, takes place in multi-segment magnetrons. It was studied by Posthumous, McPetrie, Megaw, Hervey and others, whilst Kilgore had constructed a two-segment magnetron in which a massive line circuit was part of the valve. The construction in 1940, by Randall and Boot, of a multi-segment magnetron in which the segments were part of high-Q, cavity

resonators and improvements introduced by Sayers, resulted in a greatly improved output and efficiency and made this mode of operation of the magnetron by far the most important.

For the sake of brevity, the resonance mode of operation will only be discussed in connection with the cavity magnetron but it should be understood that this type of oscillation can be produced in any multi-segment arrangement.

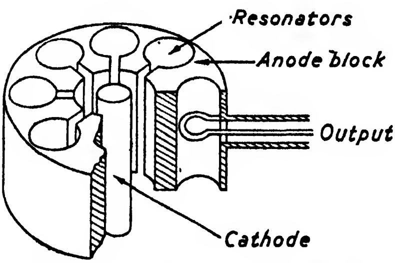


FIG. 292. Construction of Cavity Magnetron.

A sketch showing the main features of a modern magnetron is given in Fig. 292. The anode is a block of copper in which the cavity resonators are machined. It will be seen that these are of such a shape that the electric field is mainly concentrated across the gaps and the cylindrical part forms the main inductance. The resonant frequency is not greatly dependent upon the axial length of the cavities. The electric field will, of course, fringe out into the cathode-

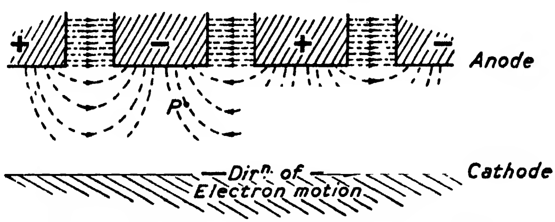


FIG. 293. Oscillating Electric—Field Distribution in Cavity Magnetron.

anode space and it is this field which interacts with the moving electrons, to maintain oscillations.

The ring of cavities can oscillate in a number of different modes, with different phase angles between the oscillations in adjacent cavities, though the total phase shift round the ring must be 2π, or a multiple thereof. The most satisfactory mode is when the phase shift between adjacent segments is π, known as the π-mode.

For this mode the electric field will be distributed somewhat as shown in Fig. 293, this field being actually superimposed on

the radial field due to the D.C. voltage between anode and cathode. The cylindrical valve is shown developed, for simplicity. During one half-cycle the magnitude of this field will grow from zero to a maximum and fall again, but the pattern will be the same. During the next half-cycle it will, of course, be reversed in direction.

Such an alternating field can be resolved into two rotating fields travelling in opposite directions. We have already seen (page 166) that two waves travelling in opposite directions (on a line, for example) form a purely stationary wave if they are of the same magnitude, frequency and waveform. The converse is equally true. We have in the present case a R.F. field which is stationary in space and varying sinusoidally with time. We can therefore imagine this field as made up of two constant-amplitude, sinusoidal fields, travelling in opposite directions at such a speed that they pass from one gap to the next in half a cycle. Some of our readers may be familiar with this concept in connection with single-phase induction motors and other machines.

We have so far confined our attention to the resonators and to the R.F. field distribution. We will now consider the interaction with electrons. Suppose that we adjust the magnetic field and anode voltage to values beyond cut-off, so that, in the absence of any oscillations, there would be no anode current and the electrons leaving the cathode would be following cycloidal paths (between plane electrodes), returning to the cathode.

Let us now further adjust the magnetic field and anode voltage so that the tangential velocity of the electrons is about the same as that of the travelling fields. Only the travelling field which is going in the same direction as the electrons need be considered, as the field going in the opposite direction will be accelerating and retarding the electrons in rapid succession and no nett energy will be either given up or absorbed by it.

An electron which gets into the field when it is tending to retard the electron will yield up energy to the field and will continue to do this as it passes a number of segments. Since it is slowing down, the force on the electrons due to the magnetic field is reduced, the curvature of its path decreases and it eventually reaches the anode.

An electron which gets into the field when this is accelerating it will absorb energy from the field. The curvature of its path increases, however, and it quickly returns to the cathode.

For a certain value of magnetic field and anode voltage, there would only be one radius at which the velocity of the electrons would be the same as the rotating field. It would appear, therefore, that most of the electrons would be rather ineffective, leading to a low efficiency. Further, the electrons would tend to separate, due to repulsion between them, so that there would be a tendency for the cloud of electrons travelling round in the most favourable position for oscillation to get dispersed.

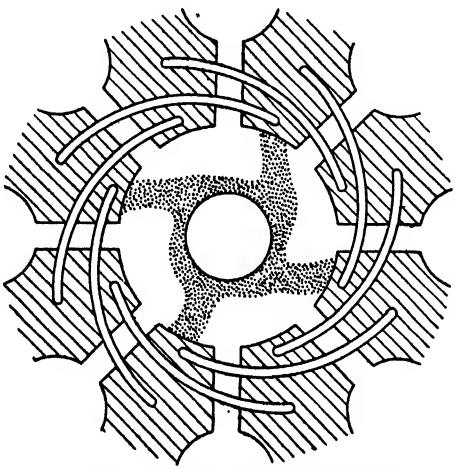


Fig. 294. Electron Motion in Cavity Magnetron.

It is here that the radial component of the R.F. field helps. Consider an electron which is at the point P (Fig. 293) in relation to the electric field, the tangential component of which is opposing the electron's motion, so that it is in a favourable position to yield up energy. If it now begins to drop back it finds itself in a stronger radial electric field, which increases its translational velocity (see equation (9)) and tends to bring it back into the most favourable position again.

Hence the electrons tend to crowd into the rotating positions where the tangential R.F. field has its maximum retarding value and so the clouds of electrons form "spokes," rotating at the frequency of the field, as sketched in Fig. 294.

It should be understood that it is the static field produced by the anode voltage which supplies the electron energy, but it is the R.F. field produced by the cavity resonators which diverts the electron energy into paths in which it can maintain R.F. oscillations.

A great deal of theoretical work on the magnetron has been done during the war years, mainly by groups working under Hartree and E. C. Stoner. This has included the calculation of actual electron paths, under oscillating conditions, and allowing for space-charge effects by a method of successive approximations. These calculations establish the general correctness of the ideas set forth above.

Comparing the resonance oscillations with the electronic, it will be seen that for the electronic type the field is adjusted to approximately cut-off, whereas for the resonance oscillation it is considerably above. The frequency of the electronic type depends upon the circular component of the electron velocity, whilst the resonance type depends upon the translational component. In the electronic type there is a difficulty in getting rid of the electrons when they have ceased to be useful and the efficiency is low. With the resonance type, favourable electrons are likely to remain so during the whole time which they spend in the cathode-anode space, and the efficiency can be high.

A very rough estimate of the relationship between the oscillation frequency B and E will be made.

Let r_c be the cathode radius and r_a the inner radius of the anode block, whilst p is the number of pairs of segments. We will assume the electric field in the cathode/anode space due to the D.C. anode voltage E to be uniform and that the electron velocity is entirely due to E.

Then $F = \dfrac{E}{r_a - r_c}$ and the translational velocity of the electrons is $\dfrac{E}{B\,(r_a - r_c)}$ (see equation (9)).

If we assume the electrons to be travelling round a mean path halfway between cathode and anode, then their angular velocity

$$\omega = \frac{E}{B(r_a - r_c)\,(r_a + r_c)\frac{1}{2}} = \frac{2E}{B(r_a^2 - r_c^2)} \text{ radians/sec.} \qquad (24)$$

The rotating field due to the oscillations in the cavities will travel past two segments in one cycle (for the π mode), that is, in $\dfrac{1}{f}$ secs.

Hence the time for one revolution is p/f and the angular velocity is $\frac{2\pi f}{p}$ radians/sec.

This must be equal to the angular velocity of the electrons, so that

$$\frac{2\pi f}{p} = \frac{2E}{B(r_a^2 - r_c^2)}$$

or

$$f = \frac{pE \times 10^{-6}}{\pi B(r_a^2 - r_c^2)} \text{ Mc/s} \quad . \quad . \quad . \quad (25)$$

If the number and resonant frequency of the cavities is fixed and B is also fixed, then the correct value of anode voltage is evidently given by

$$E = \frac{\pi B f}{p}(r_a^2 - r_c^2)10^6 \text{ volts } (f \text{ in Mc/s}) \quad . \quad (26)$$

For example, if $f = 10{,}000$ Mc/s, $B = 3{,}000$ gauss (0·3 webers/m²), $r_a = 0{\cdot}4$ cm, $r_c = 0{\cdot}2$ cm and $p = 6$, then E should be 18·8 kV.

Description of the Cavity Magnetron

Reference has already been made to Fig. 292 in order to explain an arrangement in which resonance oscillations are possible, but the cavity magnetron merits further description in view of its importance.

When used in a radar transmitter the magnetron may be called upon to give a pulse of oscillations lasting for, say, 2μ secs., at a repetition rate of 1,000 c/s. The valve is therefore only oscillating during 1/500th of the time and the peak power is 500 times the mean.

It has already been mentioned that the ring of cavities can oscillate in a number of ways. The ring is not symmetrical electrically because the output loads one of the cavities. It was found that the field pattern for all modes other than the π mode could take up either of two positions relative to the coupling loops, these alternatives being termed "doublets." A small change of conditions can cause a jump from one doublet to the other and this produces a small change in wavelength and a marked change in output. This form of instability was very troublesome in the earlier cavity magnetrons.

It was greatly reduced by strapping the segments together, one arrangement of strapping, by means of wires, being shown in Fig. 294. It will be seen that these wires connect together segments which should be at the same potential if the magnetron is oscillating in the π mode. Oscillations in the other modes are still possible but considerable currents then flow in the straps, so that their inductance is important and has the effect of making the frequencies of the other modes much higher than that of the π mode. In consequence, as the anode voltage is raised, the π mode will be excited first and mode-jumping will not occur for a wide range of operating conditions.

It will be seen that the most usual way to extract power from the magnetron is to place a small coupling loop in one of the cavities. Any lines or waveguides coupled to this must be carefully matched, if the frequency produced is to be stable. Any line carrying a large stationary wave will exhibit large changes in input impedance for small changes in frequency and, in consequence, oscillation at more than one frequency may be possible.

When considering the resonance oscillation we saw that the "unfavourable" electrons are quickly returned to the cathode, so that it is subjected to electron bombardment. As a result, there is secondary emission and the cathode temperature is also raised.

This secondary emission is one of the reasons why the magnetron has been such a successful pulse transmitter for radar. Under pulse condition san anode current of 40 A can be obtained from a cathode which is only capable of giving a primary emission of a few milliamperes. This bombardment of the cathode governs, however, the mean power that can be handled if a reasonable life is to be obtained.

The early types of magnetron all had thin, directly-heated tungsten filaments and it was generally believed that the cathode had to be of small diameter and that oxide-coated cathodes would not have a satisfactory life. Later work showed, however, that a large-diameter, oxide-coated cathode was entirely suitable and that, under pulse conditions, very large anode currents could be obtained, as already mentioned.

Early experimental magnetrons usually employed electromagnets so that the field could be varied, but the modern types

use permanent magnets and much work has been done on magnet design so as to produce the required field, which may be about 1,500 lines per cm^2 in the rather long gap (say 3 cm) occupied by the valve itself.

The cavity magnetron has been mainly developed to work around two frequencies, 3,000 Mc/s (10 cm) and 10,000 Mc/s (3 cm). The 10-cm magnetron was first developed and types

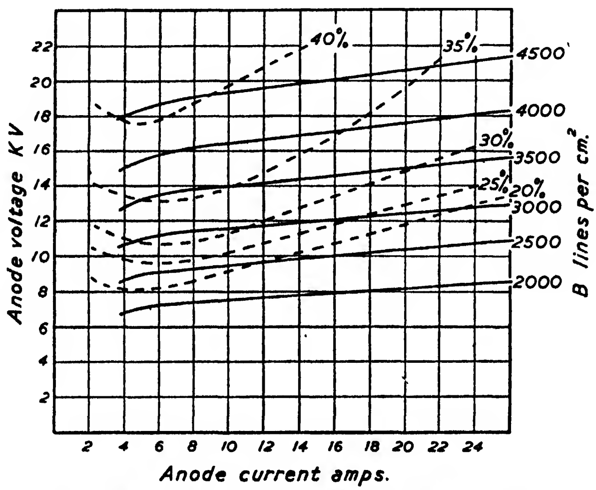


FIG. 295. Characteristics of a Cavity Magnetron. (14 Segments, 3·1 cm, Peak Power 100kW.)

produced which are capable of giving a peak power, on a short pulse, of one mega-watt, with an efficiency of 50%.

A great many difficulties were met with in producing 3 cm designs. The resonant cavities are smaller and require more accurate workmanship. The anode and cathode diameters are reduced and the number of segments increased in order that the necessary electron velocity may be obtained at about the same anode voltage as is used for the 10-cm types. The smaller anode diameter means that a larger magnetic field must be used to get well beyond cut-off, and values of 5,000 have been used. This is usually provided by two permanent magnets " in parallel " and it is necessary to have iron end-plates in the valve itself.

Designs have been produced in which the valves, magnets,

coupling and a short length of waveguide are all assembled in the test room and sent out as a unit.

The magnetrons so far described have been essentially for fixed frequency operation. The frequency can be adjusted during manufacture by bending the straps, but the frequency of the complete valve is only varied over a very small range by the external circuits—in fact, this change must be small for satisfactory operation.

Tunable magnetrons have, however, been produced. The usual arrangement is to have small pins projecting into each of the cylindrical portions of the cavities. These pins are mounted on a plate which is moved by an external control, through a vacuum bellows. Thus the amount by which the pins project into the cavities can be varied, with a consequent change in their resonant frequencies. By this means about a 10% change in operating frequency can be made.

Cavity magnetrons have also been produced for communication work. In this case, even if pulse modulation is being employed, the magnetron will be giving an output for a much greater proportion of the time and it becomes necessary to reduce the anode voltage in order that the mean power may be kept within safe limits. The electron velocity is lower and, therefore, the distance between one segment and the next must be reduced, either by reducing the inner diameter of the anode block or by increasing the number of segments. The latter is the more practical solution.

Dynatron Oscillations

This type of oscillation is only possible with a split-anode type of valve, and does not depend upon electron transit-time at all. Consider a 2-segment magnetron valve. If we connect the two segments together and plot a characteristic of anode current-magnetisation, we obtain, of course, a curve similar to that shown in Fig. 287. If, however, we separate the segments, and, for different field values, vary the voltages to each segment, we obtain a family of curves as shown in Fig. 296, from which it is seen that for the range of values of B beyond " cut-off " the segment at the lower voltage takes more current than the other. The reason for this can best be shown by considering Fig. 297 which shows the electric field when the segments are

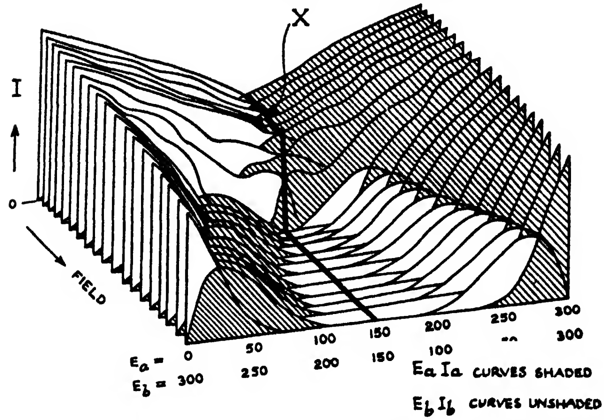


FIG. 296. Static Characteristics of Split-anode Magnetron.

at different potentials. In Fig. 297, p gives the direction of force due to the electric field, f that due to the magnetic field, which is always at right angles to the direction of motion of the electron, and the full line sketches the probable path of an electron which would have missed the anode and returned to the filament if both segments had been at the same potential E. It will be observed that the potential difference $2E_1$ between the segment distorts the field considerably near the gap and as a result the electric field has a component retarding the motion of the electron at this point in its path. This reduction of velocity decreases the magnetic force on the electron and reduces its curvature of path so that it arrives on the lower-potential segment.

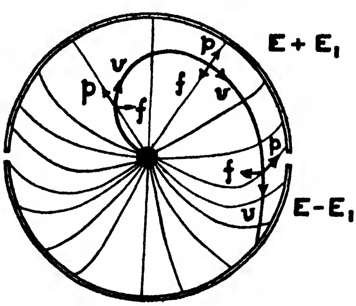


FIG. 297. Showing Electron-path Terminating at Segment having Lower Potential.

Kilgore has plotted the equipotential lines for a 2-segment magnetron when there is a considerable P.D. between the segments and shows that in this case the electron path may be a spiral, the electron arriving on the lower-potential segment, as in the case pictured above. By placing a small amount of argon in the envelope of the magnetron Kilgore was able

to photograph the electron paths due to the ionisation produced.

We will now consider the action of such a valve assuming that between the segments we place an *LC* circuit whose frequency is low compared with the electron transit-time. For the present we can therefore consider the circuit as repre-

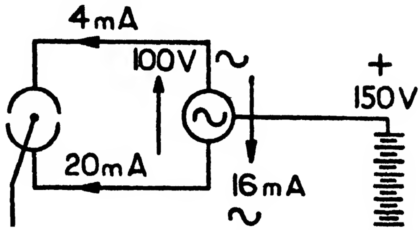


FIG. 298. Illustrating Dynatron Oscillation.

senting an alternator of zero internal resistance, giving, say, a peak voltage of 100, as indicated in Fig. 298.

If it is assumed the field is set to beyond " cut-off " as shown by the most forward curve of the series of curves in Fig. 296, then at the peak condition of A.C., segment " *a* " at 50 volts is taking 20 mA, and segment " *b* " at 150 volts is taking 4 mA. Thus, although the alternator is producing a

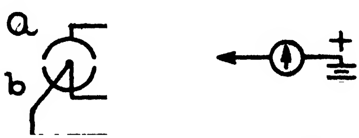


FIG. 299. Dynatron Circuit for Magnetron.

voltage of 100, it is being driven as a motor, since the current flow is in the opposite direction to the generated voltage.

It is now an easy step in the discussion to replace the alternator by a tuned circuit (Fig. 299) which, when it is oscillating, will produce an alternating potential difference between the segments. If the " motoring " effect previously mentioned is sufficient, so that the magnetron gives energy to the circuit, then oscillations will be maintained, the energy coming from the anode D.C. supply.

Of course, another way of considering the production of these dynatron oscillations is to use the negative-resistance concept. Since the voltage between the segments is in the same direction as the current which appears to flow between the segments, instead of being opposed as in an ordinary resistance, we regard the impedance between the segments as being a negative resistance which can cancel the positive resistance of the circuit placed between the segments, and oscillations can thus occur.

This type of oscillation has evidently nothing to do with electron transit time, since it is due to a property which is apparent in the static characteristics, and the frequency is determined by the circuit connected between the segments.

The conversion efficiency of the dynatron oscillation can be high, values of 50% being obtainable at 100 Mc/s, the efficiency falling rapidly as the transit-time frequency is approached.

The adjustments of this oscillator are very simple, as it is only necessary to set the circuit to the frequency required, adjust the magnetic field to an approximate value and bring up the anode voltage until oscillations are obtained (as indicated by a sudden rise of anode current). Alternatively, the anode voltage may be set to full value right away and the field varied, but always from a high value to a lower one, until oscillations are obtained. Starting from a low field and increasing may overload the valve by causing excessive anode loss.

A great deal of work has been done on magnetrons to produce dynatron oscillations and we have seen that the efficiency can be high, but such valves have found very little application in communications. This is mainly due to the difficulty of getting linear modulation, or of producing amplitude modulation without excessive frequency modulation. A great number of special circuits and valves have been proposed but, now that reaction triode-oscillators have been made to work at much higher frequencies than was previously thought possible, it seems unlikely that the dynatron oscillation of the magnetron will be employed.

Electron-Deflection Oscillators

The extensive development in recent years of the cathode-ray oscillograph and other devices in which an electron beam is

deflected has led to attempts to construct amplifiers and oscillators in which a beam of electrons is deflected by the input voltage instead of varying the number of electrons in a stream, as in the usual type of valve.

The essentials of such an arrangement are shown in Fig. 300, which may be a high-frequency amplifier if the deflecting plates are supplied from an external source, or a self-oscillator if their voltage is derived by coupling back from the output circuit. Evidently, the alternating voltage between the deflecting plates will result in the electron beam " switching " from one anode to the other, thereby giving impulses to the resonant output circuit.

Arrangements of this kind have suffered from the defect that, if the beam was made long enough so that the switching action took place with a low voltage between the deflecting plates, then the beam had a very high resistance and it was therefore difficult to obtain appreciable power and difficult to design an output circuit of sufficient resistance. In other words the mutual conductance of the arrangement compared unfavourably with that of the ordinary valve.

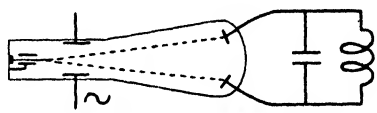


Fig. 300. Illustrating Electron-deflection Oscillator.

We have seen that the input impedance of ordinary valves decreases greatly at high frequencies and that this is a difficulty in the use of such valves. If we deflect the electron beam, instead of varying its density, then there is always the same number of electrons between the control electrodes and the transit-time effects do not occur in the same way. The input impedance can therefore be much higher than in an ordinary valve.

Such oscillators and amplifiers have been employed to a very limited extent.

Velocity-Modulated Beams

A second alternative to the conventional method of controlling an electron stream by varying the number of electrons is to vary their velocity, and such an electron stream is said to be " velocity-modulated." Control of modulation can remain

efficient and the control electrode of high impedance, even at extremely high frequencies, in contrast to the behaviour of the conventional control-grid.

Consider the arrangement shown in Fig. 301. Electrons are emitted from the cathode and formed into a beam by the focussing electrode, usually termed the "grid." If initial velocities of electron emission are neglected, all the electrons will be travelling at the same velocity as they approach the grids *GG*. V_{DC} is much larger than V, so that the electron velocity is mainly decided by V_{DC} and the time taken for the electrons to pass through *GG* is a small fraction of $2\pi/\omega$.

Electrons which enter the space between the grids when the

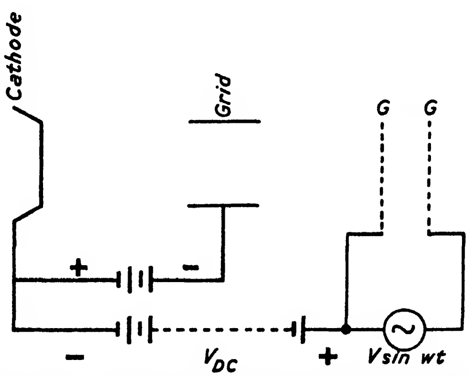


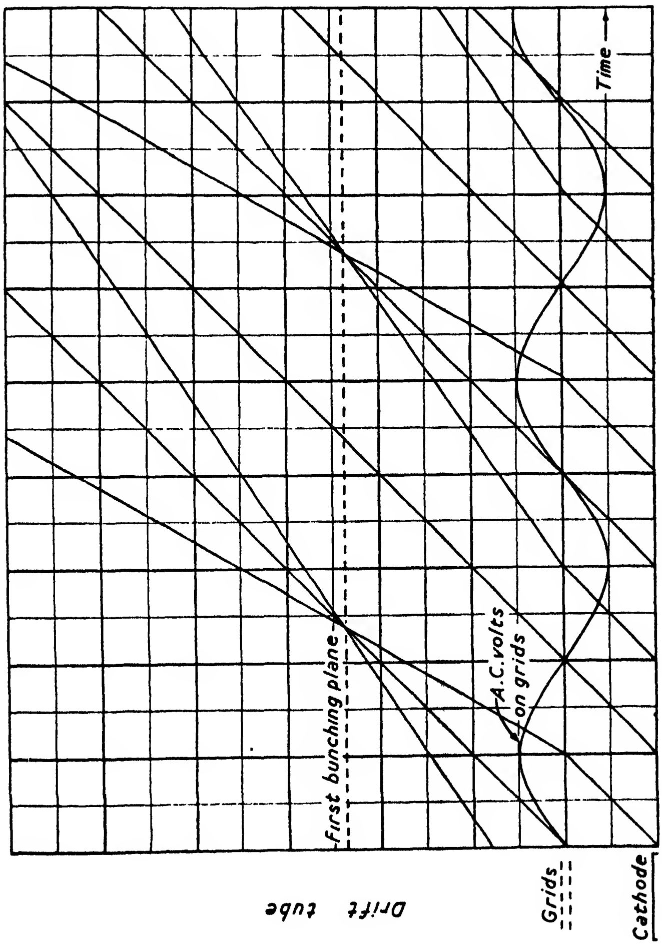
FIG. 301. Velocity-modulation of Electron Beam.

left-hand grid is positive with respect to the right-hand one will be retarded by the electric field between the grids and will, therefore, travel on into the space beyond the grids (which is all at the potential V_{DC}) at a reduced velocity. Those which arrive when $V \sin \omega t$ has reversed in polarity will evidently be accelerated and will proceed at increased velocity whilst some electrons will pass through when $V \sin \omega t$ is near zero and will be practically unaffected. Thus the voltage $V \sin \omega t$ between the grids has produced a velocity-modulated beam.

If now we allow the electrons in such a velocity-modulated beam to drift along a tube after passing the control-electrode system, it is clear that the faster ones will catch up on the slower ones and what was formerly a continuous stream of electrons will become sorted into groups so that regions where the elec-

trons are closely packed and where they are rare will be travelling down the tube.

Conditions in the drift tube can be studied graphically by the diagram shown in Fig. 302. Each line shows the position of a



typical electron at various times. The steeper the line, the faster the electron is travelling and the electron velocity is shown related to $V \sin \omega t$.

It will be seen that there are points where the number of electrons passing varies widely with time, that is, the drift of

the velocity-modulated electron beam has produced intensity modulation, which is more pronounced at certain cross-sections, termed " bunching planes."

If more elaborate space-time diagrams are drawn,[14] then the actual variation of electron density in time and space can be derived. By the courtesy of Mr. D. M. Tombs, Fig. 303 shows a model made by him which gives the relation between electron density, distance along the drift tube and time. It will be noted that the variation with time at the bunching planes is very far from sinusoidal.

We have assumed that the time taken for an electron to pass

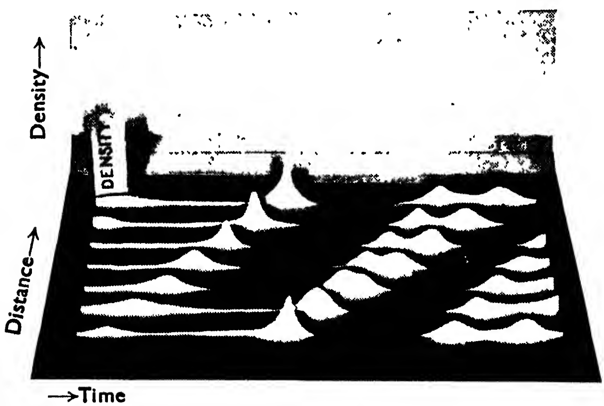


FIG. 303. Model showing Electron "Bunching".

between GG is a quite negligible fraction of a cycle of $V \sin \omega t$ and this is actually the condition for most efficient bunching. In this case the damping due to the electron stream, on any circuit connected to GG, would be negligible.

If the transit time is not negligible, then the accelerated electrons are less time between GG than are the retarded ones and, in consequence, the accelerated electrons draw more energy from GG than the retarded electrons give back.

In actual valves working at, say, 3,000 Mc/s, the transit time between the grids cannot be negligible because the spacing between them would be impracticably small. The damping imposed, however, is very small compared with that imposed upon a single grid which is modulating the intensity of the electron stream.

The Klystron

This was developed about 1939 by the brothers R. H. and S. F. Varian [11] and a diagram of an early klystron oscillator is given in Fig. 304. A cathode and focussing electrode (usually termed the grid) produces an electron beam which passes through the two grids of the " buncher."

These grids are part of a cavity resonator, termed a " rhumbatron," [12] having a cross-section of dumb-bell form. If an

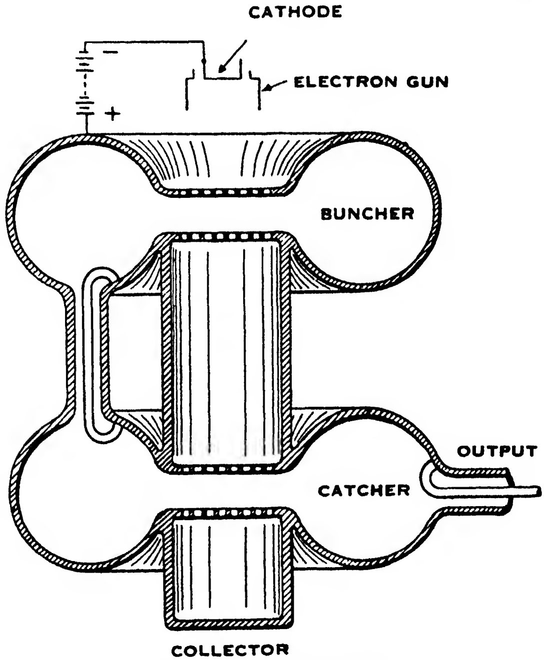


Fig. 304. The Klystron.

oscillation is produced in the buncher, there will be an alternating electric field between the grids, which produces velocity modulation of the electron beam in the way previously discussed.

By adjusting the D.C. voltage, and, therefore, the velocity of the electrons, we can arrange that the catcher is at a bunching plane, so that the electrons pass between the catcher grids in bunches.

Suppose that the catcher is oscillating and ignore for the time

being the concentric line between catcher and buncher. If the phase of the oscillation is such that when the electric field between the grids is retarding the electrons, the bunched electrons are passing through the catcher, whilst few electrons are passing when the field is accelerating them, the electron beam will be yielding up energy to the catcher and a large conversion of energy from D.C. to A.C. is thus possible. For high efficiency the catcher should have a high Q-value, so that a powerful electric field is produced across its neck and hence the electrons are retarded to nearly zero voltage across the catcher.

Evidently, if the oscillation of the buncher is maintained by feeding back a small portion of the oscillating energy of the catcher by way of a concentric line of the right length to give the right phase, we have a self-oscillating system.

After having designed the klystron so that the spacing between buncher and catcher and also the length of the concentric line is suitable, the operating adjustments will be two—tuning the rhumbatrons to the same frequency and adjusting the D.C. voltage so as to produce a bunching plane at the catcher.

It is necessary for the rhumbatron to be somewhat of the shape shown in order that electrons accelerated by convenient voltages may pass through it in less than a half-cycle. If a sphere had been used, for example (with the electron stream passing through its diameter), then its wavelength is only slightly greater than its diameter, so that, even if the electron stream could have the velocity of light, it could not get through in a half-cycle of the oscillation. With the shapes of rhumbatron used, however, a voltage of about 3,000 volts is sufficient to produce the correct velocity, about 1/10th that of light, for the correct phasing of the beam.

The klystron oscillator was the first to employ high-Q cavity resonators as an integral part of the oscillator. The original rhumbatrons were difficult to construct and modern types use simpler shapes. It is not an easy matter to provide adjustable tuning and most of the two-resonator klystrons are made for a fixed frequency. It is usual to dispense with grids where the electron beam passes through the resonators. If the gaps are correctly proportioned it is possible to get the proper

interchange of energy between electron beam and resonator without grids.

The original klystrons worked at about 3,000 Mc/s and gave an output of 300 W, and klystrons of similar rating have been used as C.W. oscillators during the war, their efficiency being about 10%. The klystron has been used only to a limited extent to produce the high-power pulses required for radar, because it was soon found that the magnetron was superior for this purpose. However, a klystron was produced giving a peak power of 25 kW on about 3,000 Mc/s.

Modulation can be applied to the "grid," or focussing electrode, of a klystron, producing the same effect as in a television cathode-ray tube. The number of electrons forming the beam is changed and this, of course, varies the energy supplied by the electron beam to the catcher.

The frequency stability of the klystron is better than that of most electron oscillators and the application of modulation to the grid introduces very little frequency modulation.

The klystron can evidently be used as an amplifier if the buncher is supplied from an external source instead of from the catcher. A good deal of work on klystron amplifiers has been done and power gains of 100 times have been obtained. Unfortunately, the klystron amplifier is not (in its present development) an answer to the problem of providing signal-frequency amplification in a superheterodyne receiver (see page 571) so as to improve the signal/noise ratio. This is because it produces much more noise than an ordinary type of valve. It is not clear at present whether this can be reduced sufficiently to make the klystron a useful amplifier for small signals.

When the klystron is arranged as an amplifier it can be used as a frequency-multiplier by tuning the catcher to a harmonic of the buncher frequency. The efficiency is still good because the intensity variation of the beam at the bunching planes is far from sinusoidal.

The Reflector Klystron

This is a modified type of klystron in which there is only one resonator. The electron beam passes through a resonator, where it is velocity-modulated (if the arrangement is oscillating) and is then turned back by a reflecting electrode, so that it

passes again through the resonator. During the passage from resonator to reflector and back again, the beam will have become bunched. If the bunches pass through the resonator when its electric field is retarding the electrons, energy is evidently given up to the resonator. If the energy given back is greater than the losses in the resonator (including the energy required for bunching) then the arrangement will oscillate.

Klystrons with one resonator are simpler to construct and to adjust and were greatly used during the war as local oscillators for superheterodyne receivers working on frequencies such as 3,000 Mc/s.

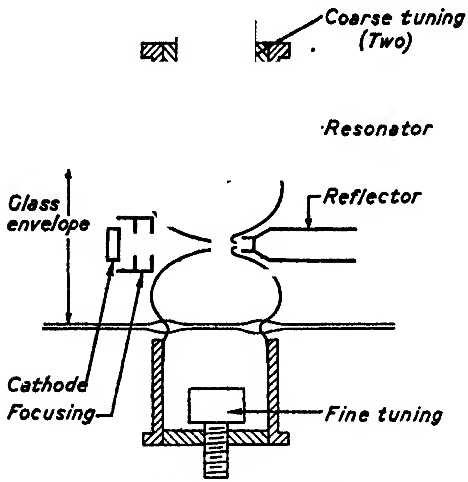


FIG. 305. Reflector Klystron.

There are at least three different ways of arranging the reflection. If we place a sufficient negative voltage on the reflector and shape it so that the retarding field is approximately uniform for some distance in front of it, then the faster electrons will go nearer to it than the slow ones and, having had farther to travel, may arrive back at the resonator later than the slow electrons.

If, however, we shape the reflector so that its field is mainly concentrated just in front of it, then all the electrons will travel about the same distance and the faster ones will, of course, arrive back first.

A third possibility is to allow the electrons to reach the reflector and produce secondary emission there. If the reflector is only slightly positive, then the number of secondary electrons emitted will be greatly dependent upon the velocity of the primary ones and hence bunches of electrons will pass back to the resonator. In this arrangement bunching will be seen to be only partly dependent upon "drift."

Fig. 305 shows the construction of a reflector klystron

suitable for 3,000 Mc/s, in which the reflector works under the condition first discussed above. It will be seen that the glass envelope goes through the resonator and there is, in consequence, some dielectric loss in it. This construction, however, makes it easy to arrange for tuning (by screwing plungers into the cavity) and coupling to the output (by a loop and concentric line, not shown in the figure). The plungers provide a tuning range of approximately 8% and a small adjustment of frequency can also be made by altering the reflector voltage.

In any velocity-modulated arrangement oscillations can only take place over a range of voltages, for which the drift time gives the correct phase relationships. In this type of reflector klystron, however, the adjustments are not critical. For example, a test showed that oscillations occurred with any anode voltage between 1,000 and 2,000 volts (with fixed reflector voltage), the efficiency being a maximum (3·5%) at 1,500 V. The reflector voltage could be varied (with fixed anode voltage) from about − 180 V to − 330 V with maximum output at − 260 V, the output being more dependent upon correct reflector voltage than anode voltage.

It will be gathered that the efficiency of these reflector klystrons is not high but they are very convenient where outputs such as 200 mW are required.

A reflector klystron has been developed for use in measuring equipment, in which the valve fits into a concentric line forming the resonator. The working length of the line can be varied by a non-contact type of piston and a tuning range obtained of, for example, 7–14 cm.

If a wide frequency-range is to be obtained with any type of electron oscillator it will be necessary to vary the operating voltages as well as the resonator tuning. In this oscillator the tuning control and the potentiometer controlling the reflector voltage are ganged together so that correct conditions are preserved over the whole tuning range.

Concentric-Line, Velocity-Modulated Oscillator

This belongs to a class of oscillator in which there is only one resonator, but it has two gaps, one of which acts as a buncher and the other as catcher. The beam does not turn back on its tracks.

The electron beam passes transversely across a concentric line, as seen in Fig. 306, and the buncher and catcher gaps are between the inner and outer conductors. If the shape of the line was not modified, where the beam passes through, then the gaps would be too long. This could be got over by increasing the diameter of the inner conductor but this would make the capacitance high and reduce the resistance which the resonator puts across the gaps. The line is therefore modified in the manner shown.

In the two-resonator klystron we can evidently adjust the feed-back by altering the coupling loops but in the concentric-

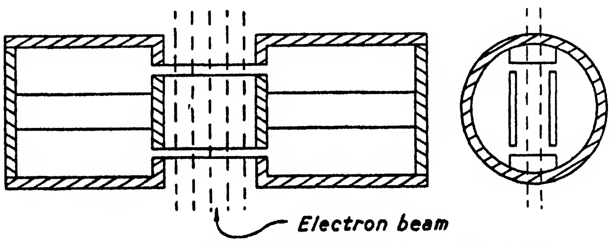


FIG. 306. Concentric-line Velocity-modulated Oscillator.

line oscillator the coupling is essentially unity, because the field strength in both gaps will be the same.

Practical valves of this type employ a permanent magnet which produces a field parallel to the electron beam. This is to prevent spreading of the beam due to repulsion between the electrons.

The portion of concentric line within the valve can be coupled to a line or resonant cavity and by altering the dimensions of this a wide tuning range, such as 2 : 1, can be obtained. The anode voltage needs to be varied in order that the transit time may remain correct.

Travelling-Wave Valve

We have seen that, in the klystron, it is desirable that the time spent by an electron in traversing the electric field at the buncher should be a small part of the cycle but at the highest frequencies this would lead to impracticably small spacings. An alternative is to launch both an electromagnetic wave and an electron beam down a tube, with very nearly the same

velocity. An electron therefore stays in a field having the same phase for a considerable time. The interaction between the wave and the beam can be such that one component of the wave absorbs energy from the beam and increases in amplitude. Thus the travelling wave valve can be used as an amplifier or, if the input is coupled back to the input, it can be used as an oscillator.

The wave should have a component of electric field in the direction of propagation (and therefore along the electron beam) in order that some of the electrons may be retarded and yield up energy to the wave. It is also necessary that the velocity of the wave should be about one-tenth that of light,

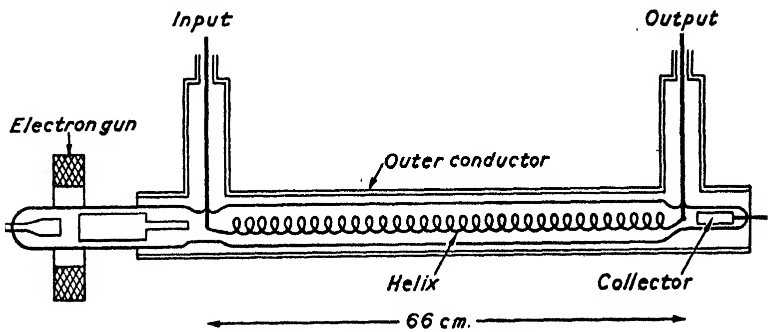


FIG. 307. Travelling-wave Valve.

otherwise excessive accelerating voltages would be required. These conditions are satisfied by a concentric line in which the central conductor is a helix, which should be several wavelengths long, in order that the effects may be cumulative.

In the arrangement shown in Fig. 307 (due to Kompfner) an electron gun produced a beam current of about 50μA, for an anode potential of 1,800 V, the power amplification being about 14 times at a frequency of 3,000 Mc/s. There was a considerable improvement in signal/noise ratio, so that it is possible that these valves may be developed for use as amplifiers in front of the mixer in centimetric-wave receivers.

Since the velocity of the electron beam must be very nearly the same as that of the wave (actually slightly greater) for appreciable amplification to take place, it is to be expected that the anode voltage would be critical. In a typical case a

3–4% reduction in anode voltage below the optimum value resulted in the valve attenuating. As there are no sharply resonant circuits, the amplification obtainable is the same over a wide frequency band.

Similar valves, giving somewhat larger outputs, have been constructed by Pierce and Field, for use as power amplifiers.

CHAPTER XIII

MODULATION CIRCUITS

In the chapter dealing with modulation theory it was shown that the problem is to vary either the amplitude, frequency or phase of a radio-frequency carrier by the modulating signal. It was also seen that the modulation must vary only one of these parameters, any accidental variation of the others causing distortion and, possibly, loss of modulation.

In the case of amplitude modulation, the modulation wave-shape appears as the envelope of the carrier and the carrier amplitude changes are carried right through the system, so that at any stage subsequent to the modulator the envelope of the R.F. wave should follow the modulation waveform. With phase or frequency modulation, however, the modulation creates a change of phase or frequency from a given datum, but since the amplitude remains constant no such similarity exists between the modulated carrier and the modulation. It is at the receiver that the relative frequency or phase changes are translated back into amplitude changes.

Circuits for Amplitude Modulation

In the ideal case not only must the envelope of the modulated carrier be a replica of the signal wave-shape but it should reproduce changes of signal amplitude in a linear manner between the required limits of modulation depth and must be capable of doing so over the whole band of frequencies involved. This band will be limited in the case of commercial telephony, considerably wider for high-fidelity broadcasting and very much wider for television.

The degree to which the above conditions are fulfilled depends very much upon the services required of the transmitter. Thus with broadcasting, since the primary object is entertainment, a high degree of linearity is essential over a very wide frequency band, and over a widely varying depth of modulation percentage. Power efficiency is not a primary

consideration and the simplicity of circuit is of minor importance. Although broadcast transmitters have been designed in which the power is varied in sympathy with the varying degree of modulation so that at no time is a greater power used than is necessary and a high power-efficiency thereby obtained, no such refinement is in general use and most broadcast transmitters operate very inefficiently at an average modulation percentage of rather less than 50%. In the case of transmitters handling commercial telephony the requirements are different. Here the modulation-frequency band is limited as much as is possible and frequency response need not be strictly linear, but a high power-efficiency is usually required and (most important of all) the modulation depth must be kept as high as possible at all times in order to obtain the best possible signal/noise ratio at the receiver from the carrier power available (see page 635). These latter requirements may also be said to exist for most of the small, general-purpose transmitters. Except for the rather special circuits for television, modulation circuits both for broadcasting and other purposes are the same in general principle, but the greater fidelity of response in the former necessitates more elaborate and costly apparatus.

As mentioned in Chapter III, having modulated a carrier, the whole spectrum of H.F. may be radiated, or the carrier and one side-band may be suppressed, as is done with single side-band transmission. We will deal first with the general modulation circuits and since a single side-band transmitter is merely a special case involving no new principles it will be dealt with in its appropriate chapter. Certain of the methods of modulation described can be applied to oscillators or to power amplifiers although we are usually only interested in applying them to the latter. If an oscillator is amplitude modulated, for instance, an amount of frequency modulation is introduced which is larger than can be tolerated even if such modulation can be arranged to have no adverse effect upon the oscillating conditions.

Amplitude modulation may be carried out at the output of the final stage of a transmitter, in which case it is known as a high-power modulation system. In such a system the high-frequency amplifier chain is dealing with unmodulated H.F.

throughout, linearity is therefore unimportant and all stages will be operating at a high power efficiency. It is clear that in the case of high-power transmitters, since the modulator has to handle the full output power of the transmitter, considerable amplification at the signal frequency is necessary. Alternatively, modulation may be carried out at any stage prior to the final one, in which case the system is known as a low-power modulation system. In such a transmitter all stages subsequent to the modulated stage have to handle a modulated input at H.F. This means that each H.F. stage must have a linear relationship of input to output, necessitating a low power-efficiency as will be seen when such circuits are discussed. But since modulation can take place at any convenient stage at low power, a small and orthodox signal amplifier can easily be designed. Thus with high-power systems we have highly efficient H.F. amplifiers but large and inefficient modulators, whereas with the low-power system the reverse holds, and there is but little to choose between the two as regards overall efficiency. Both systems are widely used. An H.F. amplifier can be modulated by applying the modulation in either anode, grid or cathode circuit, and in the case of a pentode valve either of the grids may be used, the point of application classifying the type of modulation.

Anode Modulation

One of the simplest methods involves varying the anode supply volts to the H.F. valve in accordance with the modula-

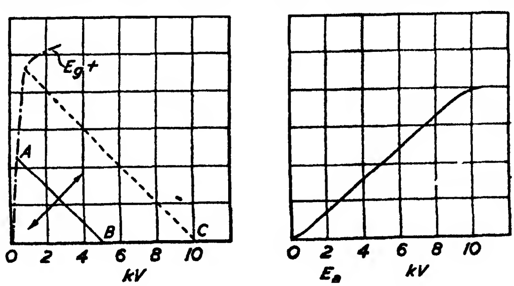


Fig. 308. Illustrating Anode Modulation.

tion, the modulator circuit being used either in shunt or series. Consider any single stage of an H.F. amplifier as shown, say

in Fig. 208, page 344, and assume the valve is being operated under Class B conditions, with an anode supply of E volts, and being driven from an input frequency of f_c into a resonant load at this carrier frequency. The load line will be roughly as indicated by ABC, Fig. 308a, and we can assume the conversion efficiency would be approximately 70%, this giving an output high frequency current I, corresponding to an average anode voltage of 5kV, Fig. 308b. If now we raise or lower the D.C. voltage, still keeping the valve under Class B conditions, the load line will slide parallel to ABC, the conversion efficiency will not materially change and in consequence

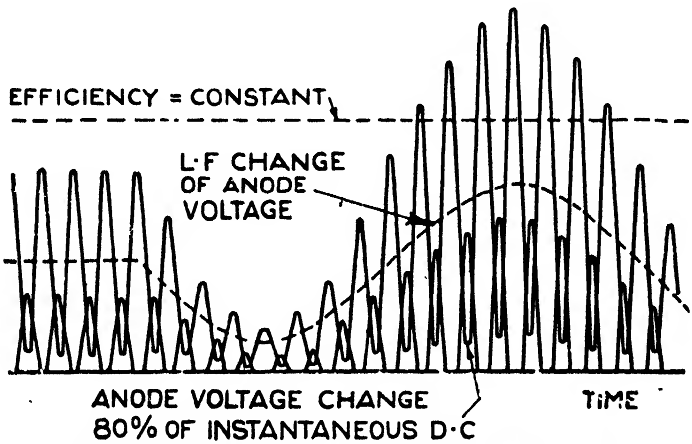


Fig. 309. Pictorial Representation of Anode Modulation.

we shall obtain a linear relationship of output H.F. current to anode voltage as shown in Fig. 308b, provided we have sufficient peak emission in the valve.

If a sinusoidal change of anode voltage is provided from a modulator system, the relationship of anode voltage and current will be as shown pictorially in Fig. 309. Observe that if the conversion efficiency remains constant, the power supplied and converted surges from high to low levels each modulation half-cycle. At the peak of modulation, the instantaneous voltage and current are doubled, hence the power is quadrupled. If the carrier is modulated 100%, then the average power rises to a value of 1·5 times the power in the carrier condition, and there will be in consequence a rise of

H.F. current during such modulation of $\sqrt{1 \cdot 5}$. This is obvious since the power due to carrier and side-bands is 1·5 the carrier power and this must be equal to $I^2_{R.M.S.} \times R$.

For other percentage modulations the rise will be correspondingly less, the relationship between modulation percentage and H.F. current (for a carrier current of unity) being shown in Fig. 310.

Conversely, the ratio $\frac{I_{mod}}{I_0}$ can be used to determine the percentage modulation, a simple method often adopted in practice. It should be pointed out that, except at high modulation levels, the measurement is not very accurate

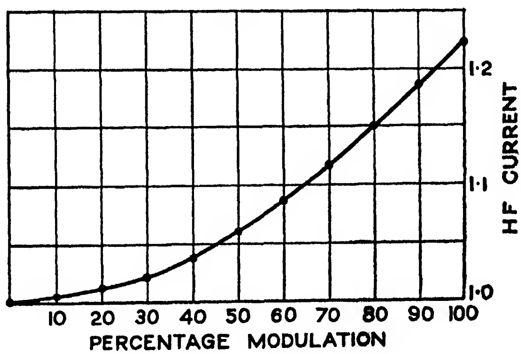


FIG. 310. Variation of R.M.S. Current with Percentage Modulation.

since there is such a very small current rise, as can be seen from Fig. 310. Further, care has to be taken with any such measurement that only a sine signal is applied, and no H.F. limitation occurs, an oscillograph check of output waveform being very desirable.

Since the valve being modulated has to handle an average of 1·5 times the carrier power during full modulation, its dissipation will also rise by the same amount, assuming the same conversion efficiency. This fact, and the greater peak emission required, has to be kept in view when considering the valve circuit. Or, conversely, for a valve designed economically for telegraph conditions we must, if it is to be used to deliver modulated H.F., reduce the carrier level by an amount which

depends upon the maximum modulation, k, required, thus:

$$\text{Modulated Power Possible} = \left\{\begin{matrix}\text{Unmodulated}\\ \text{Power}\end{matrix}\right\} \times \left(\frac{1}{1+\frac{k^2}{2}}\right).$$

Shunt Modulator and Choke

This method, due to Heising, and sometimes called the constant-current method, was one of the earliest, and is still in very common use. It employs a modulator valve in shunt with the amplifier and the supply. The circuit in its simplest form is shown in Fig. 311, where $V_aL_aC_a$ is the amplifier valve and resonant output circuit, driven from an H.F. input

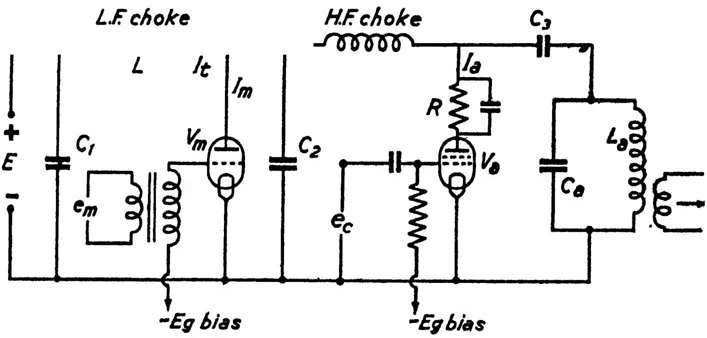


FIG. 311. Shunt Modulator.

voltage e_c and shunt fed through the H.F. choke in the usual manner. V_m is the modulator valve on whose grid is applied the signal voltage, e_m, and L is an iron-cored choke in series with the D.C. supply voltage E.

The choke circuit is more usual as the design of a satisfactory transformer for such purposes has only really become practical with the advent of the push-pull modulator such as described on page 517, an elaboration that is unnecessary except for the larger types of transmitters.

From the modulation frequency point of view we may regard the circuit as equivalent to the network of Fig. 312, where the modulator is replaced by an alternator of the signal frequency, having a peak voltage which is determined by the input signal voltage and the circuit constants (as will be shown later) across

which are two main shunt circuits and others less important. First, the comparatively low resistive load comprising the H.F. valve and associated circuit indicated by R_a, and second, the very high impedance choke L through which is fed the D.C. supply. The additional shunt condensers indicated by $C_2C_3C_a$ are all of high reactance and only influence the frequency response at the high end of the audio frequency scale, and the capacity C_1 representing that of the supply circuit in series with the choke may be regarded as a short circuit. If the choke is regarded as almost infinite impedance to L.F. (but has negligible D.C. resistance) then the alternator can be regarded as controlling the A.C. voltage and power to the H.F. valve at the signal frequency, the latter converting this into modulated H.F. as was indicated in a previous section. Since, however, the power has to be derived from the D.C. supply, the alternator being in a sense fictitious, the action is rather more complex and can best be discussed by considering the valve characteristics of the modulator in the usual manner.

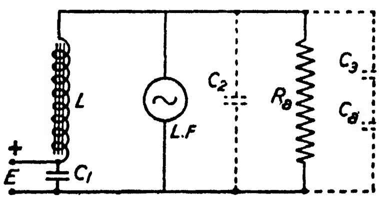


FIG. 312. Equivalent Circuit of Shunt Modulator.

Considering only the main circuits indicated above, the valve circuit simplifies to that of a modulator valve supplied from a D.C. voltage of E, whose anode circuit consists of a resistive load of value $R_a = E^2/W$, where W is the total power required at the anode of the H.F. valve and E is the D.C. voltage applied. As an example let us take the amplifier circuit designed in Chapter X, as that to be modulated.

Example. Required to modulate an H.F. amplifier from a D.C. supply volts of 5,000, having an input power to the amplifier of 1,440 watts.

$$\text{D.C. feed} = \frac{W_a}{E_a} = \frac{1{,}440}{5{,}000} = 290 \text{ mA.}$$

$$R_a = \frac{E_a}{I} = \frac{5{,}000}{290} \times 10^3 = 17{,}250 \text{ ohms.}$$

Since the modulator has to operate under Class A conditions the average modulator feed will need to be rather more than

that of the amplifier and under quiescent conditions, therefore, the power dissipated by the modulator must be somewhat greater than that delivered to the amplifier.

Fig. 313 shows the characteristics of a modulator valve suitable for the power example given above, across which has been drawn the resistive load line for $R_a = 17{,}250$ ohms. When no signal is applied the modulator grid-bias will be adjusted to give a modulator feed current I_m of 300 mA at 5,000 volts shown by the point " 0," which has to be below the maximum dissipation curve of the valve. During such quiescent condition, the feed, I_m, is approximately equal to I_a, 300 mA,

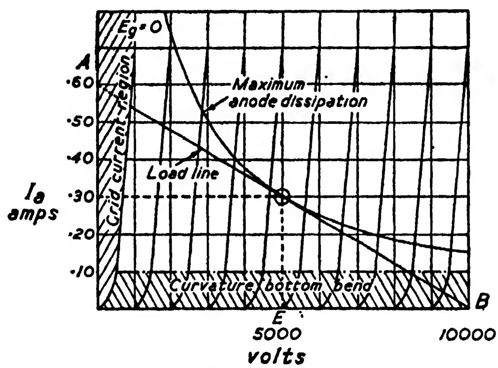


FIG. 313. Modulator Characteristics.

and thus there is a total feed, I_t, 590 mA from the supply through the choke L. This means that of the total power of 2,950 watts supplied, rather more than half (1,500) passes through the modulator as dead loss, and the remainder (1,450) goes to the amplifier, which in turn converts 70% of it into carrier power, represented by $I_o = 1{\cdot}0$ amp. in the output, the relevant figures for the circuit being shown in Table XXI, column 2.

When an L.F. signal is applied to the modulator grid there is an in-phase modulator current produced since the load is resistive, but an anti-phase change of modulator anode voltage, these EI changes being along the load line AOB.

The changing modulator current cannot be satisfied by the supply because of the choke L whose reactance is high enough

to prevent such changes, and hence any variation of modulator feed is associated with an opposite change of amplifier feed, the latter rising as I_m falls, and vice versa. Since E_m and I_m are in opposite phase, E_a and I_a are therefore in-phase, and thus during modulation the amplifier gains power at the expense of the modulator, which loses it by the same amount, the total power supplied remaining constant because of the choke. The figures in column 3 of Table XXI show those under modulation conditions for a sine signal for 80% modulation. It is of interest to observe that since the feed instruments only show average conditions, they indicate the same values throughout, and only the aerial current shows an increase from 1·0 to 1·15 as given in Fig. 310. During modulation the modulator valve will have to dissipate less than under quiescent conditions, and with the glass-envelope type the cooling of the anode during modulation is usually observable.

For the modulator to bring about 100% control, the excursion of anode volts would need to drop to zero at the current peak, and twice E at current zero, along the load line AOB. From an examination of the curves of Fig. 313, such a swing is clearly impossible. In fact since distortion will be introduced if we run into grid current, the limitation at the peak current end is the slope of the zero grid current curve. We shall also get distortion if we operate into the bottom bend characteristic, these two distortion areas being shown shaded. We can, of course, avoid the bottom bend region by raising the average

TABLE XXI

	Carrier Average	Sine Mod. 80% Average
E	5,000	5,000
E_a	5,000	5,000
E_m	5,000	5,000
I_t	590	590
I_m	300	300
I_a	290	290
W_t	2,950	2,950
W_m	1,500	1,100
W_a	1,450	1,850
I_{HF}	1·0	1·15

point "0," thus increasing the valve loss, and it is therefore the grid current curve that is the real limitation to 100% voltage control. It is evident therefore that we must use a very low impedance modulator such that E_aI_a is as steep as possible, but even with such valves a percentage modulation of 80 is usually the maximum possible with the circuit being discussed.

Where a large band of signal frequencies is involved the straight line load characteristic shown does not represent the true load on the modulator at all frequencies; for at low frequencies the reactance of L is not infinite as assumed but may be commensurate with, or even be small compared with, R_a, whereas at higher frequencies the reactance of the shunting condensers will be small. In consequence, at the top and lower end of the frequency band the load line is no longer straight but opens out into an ellipse. This means the percentage modulation will fall away at these frequencies, and unless the average feed is raised to accommodate the ellipse, distortion will result.

A consideration of the operating conditions of the modulator shows that we need a valve which will have a high, anode-loss rating compared with the amplifier, but since the peak/average current is hardly more than 2/1, the cathode will be relatively smaller than that of the amplifier. The impedance will need to be low as we have already explained and therefore a triode valve will be selected.

Chokes are suitable only for sets of low power, as the D.C. component of feed necessitates a very large iron section if saturation of core is to be avoided. This may be overcome by the use of a tapped choke, usually 1/1 ratio, where the connections are such that the modulator and amplifier feeds pass in opposite directions; thus the magnetising effects of these feeds cancel if they are equal, but any variation of feed by the modulator impresses its voltage changes on the amplifier in a similar manner to that described in the choke circuit.

Circuit for 100% Control

It is easily possible, by a small modification, to obtain a choke-modulated transmitter to give the full 100% control. Thus if a resistance, R, shunted by a condenser large enough to have

negligible reactance to the signal frequencies, is inserted in series with the amplifier anode circuit as shown in Fig. 311, this will drop the amplifier voltage by an amount I_aR, where I_a is the average feed. This means, of course, that the volts to be converted by the modulator will be lower by this amount and if as example the modulator can only handle 80% of E, we should arrange for the IR drop to be 20% of E. Such a circuit is, of course, less efficient because of the loss in the anode resistance. If the same output was needed we should, of course, have to raise the D.C. supply voltage and that applied to the modulator, so as to leave the amplifier voltage at the same value.

The Negative Feed-Back Amplifier

It is necessary to digress at this point to consider the negative feed-back amplifier because this is frequently used in the modulating circuits of broadcast transmitters.

We derived, in Chapter XI, a general expression for the gain of an amplifier with feed-back, but then confined ourselves to the case where this feed-back is positive. If β is made negative it will be seen that the gain is $\dfrac{\mu}{1 + \mu\beta}$ and will always be less than μ, and, in fact, if $\mu\beta$ is much greater than 1, then the overall gain is almost independent of μ.

It is probably simpler to indicate the difference in action between a straight amplifier, A, and a negative feed-back amplifier, B, to do the same work, by taking an example.

Let us suppose we have a 10 V signal and wish to amplify this to 1,000 V for the purpose of modulating a transmitter. The final stage of both A and B has to do the same work, and this stage introduces 10% distortion, giving an R.M.S. voltage of 100 V.

Considering A, the amplifier must be built to give a gain of 100, i.e. 20 dbs.

To provide the same output from the same input with the negative feed-back amplifier we shall need an increased gain which is dependent upon the value of β. Let us take a value of 0·009 for β. Then substitution in the equation for the overall gain shows that the amplification will need to be 1,000.

The increased amplification necessary is clearly a disadvan-

tage but this may be offset by some advantages now to be considered.

Suppose that changes in supplies cause the gain of both amplifiers to change by 10%. The output of A is clearly changed by 10%, namely to 1,100, but that of B is given by

$$E_o = \frac{1{,}100 \times 10}{1 + (0{\cdot}009 \times 1{,}100)} = 1{,}010\ V$$

so that only a 1% increase in output voltage has occurred. This stabilising property of negative feed-back is of great importance in measuring apparatus and in repeaters for telephone lines.

Let us now compare the distortion present in the outputs of A and B. For A this is 100 V. Let the value in B be D, then distortion voltage applied to input is $\cdot 009D$ and this gives $1{,}000 \times \cdot 009D$ at the output stage, to be added to 100 V produced there, so that

$$D = 100 - 9\,D$$

$$\text{or } D = 10\ V.$$

Consequently, B produces much less distortion if the final stage is the same as in A. Alternatively, if a 10% distortion is permissible, it may be simpler and cheaper to build an amplifier which in itself produces greater distortion and apply feed-back to reduce it to 10%.

If " noise " is produced in the final stage (for example " hum " due to A.C. filament heating) then an argument similar to that above will show that the output of noise from B will be 1/10th of that in A. In the case of noise or distortion produced in the input circuit of A or B, however, there will be no reduction in the output it produces in B. Noise or distortion introduced at intermediate stages in the amplifier, however, will give less output from B than A but the ratio will be less than one-tenth.

Even if β is the same for all frequencies the frequency/response curve of the amplifier will be made more uniform because we have seen that any change (including frequency) which varies μ will vary the overall gain much less.

The presence of the feed-back connection modifies the apparent internal impedance of the amplifier output, the nature of the change depending upon the circuit.

It will be appreciated that since μ and β are actually complex quantities, a given arrangement of feed-back which is negative over a range of frequencies may become positive for other frequencies. Careful design of the feed-back circuit is seen to be necessary, particularly when the amplifier has several stages.

For a fuller discussion of negative feed-back and possible circuit arrangements, the reader is referred to any modern book dealing with amplifiers. Negative feed-back is sometimes applied to a broadcast transmitter by rectifying a portion of the modulated radio-frequency output, thus extracting the modulation, and applying this to the modulation input (in the correct phase to produce negative feed-back). Any distortion or "hum" introduced by the modulator and modulated amplifiers is thereby greatly reduced.

Class B Push-Pull

By utilising two valves in push-pull and operating them at a point just above cut-off, a system is provided in which the D.C. loss in the modulators can be avoided. Thus instead of

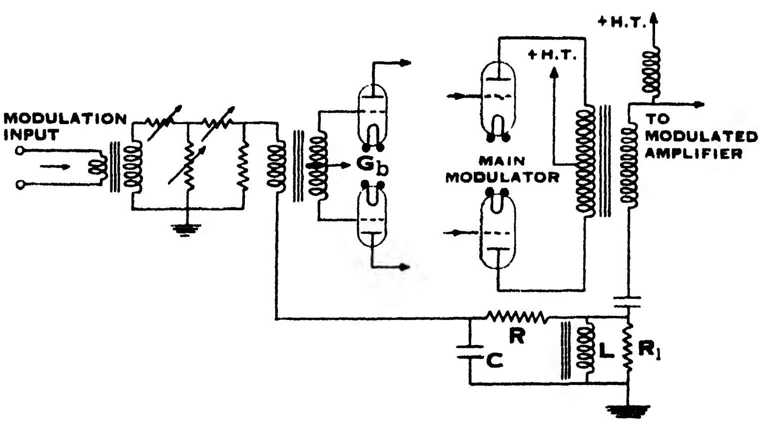


FIG. 314. Class B Push-Pull Modulator.

the modulator being connected in shunt with a choke, the two modulators are connected in push-pull across the primary of a transformer, the secondary of which is in series with the H.T. supply to the main unit to be modulated, i.e. the modu-

lated amplifier, the circuit diagram of a modulation unit being as shown in Fig. 314. To avoid passing the D.C. feed to the modulated amplifier through the transformer secondary it is usual to shunt-feed the latter through a choke as indicated. Such a circuit is often supplied with negative feed-back to correct for distortion which may be introduced by the main transformer, Fig. 314 showing a feed-back circuit of simple character.[9] In this circuit the feed-back voltage is derived from a resistance unit R_1, in series with the main transformer secondary, the voltage from which is injected back into the primary of the line-to-grid transformer of the first audio stage. To ensure stability of working and to provide a certain amount of audio frequency discrimination a resistance-capacity choke network, *RCL*, is connected across the feed-back resistance.

A trouble associated with early push-pull modulators was that due to excessive voltages being developed across the transformer during over-modulation periods. This can be prevented by the fitting of a limiting device, usually a neon tube, across a convenient modulator valve.

Anode Modulation by Series Modulator

The series modulation circuit has been developed to operate in one of two ways. In the first the modulation is applied between grid and cathode of the modulator valve, and in the second the modulation is applied in such a way that it includes the load circuit as well as the grid-cathode circuit of the modulator valve, in which case the circuit is of the type known as a " cathode-follower."

Dealing with the first case, where the modulation is applied between grid and cathode, it is immaterial whether we arrange for one side of the H.F. circuit, or one side of the modulator, to be at earth potential. If the former, we shall have the problem of feeding in the modulation at a H.F. potential above earth, whilst in the latter we have to arrange an H.F. circuit which is all at the modulating voltage above earth. Both circuits have been used, although the latter is probably a rather more simple problem, and Fig. 315 shows a series modulation circuit with modulator at earth potential.[3]

The circuit may be considered as consisting of two resistances in series, the amplifier resistance R_a being of fixed value, and

equal to $\frac{E_a}{I_a}$; and the modulator resistance R_m, which is variable. It is clear that if R_m can be varied between values of zero and infinity, the anode voltage to R_a will vary between E_b, the supply voltage, and zero, but as with the choke method of modulation, limitations of the characteristics prevent this range being achieved.

In analysing this circuit we can follow exactly the same procedure as in the choke modulation case, by setting up the

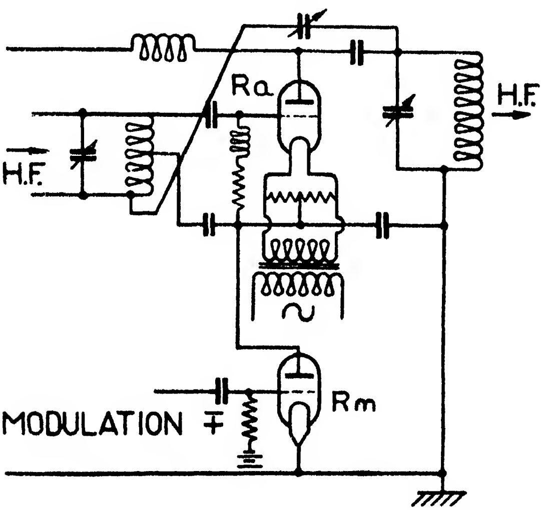


FIG. 315. Circuit for Series Modulation.

E_aI_a characteristics of the modulator valve, and laying across these the load line due to the H.F. amplifier.

As an example let us use the same modulator valve and high-frequency circuit as in the choke modulation circuit just discussed. The high-frequency circuit forms the load in series with the modulator and its value $R_a = \frac{E_a}{I_a} = \frac{5{,}000}{\cdot 30}$ giving a load line of slope AB. Then if we had an initial supply E_b of 10,000 volts (instead of 5,000) the load curve would be BOA (Fig. 313) and a bias on the modulator to the point O would determine the quiescent condition.

As before, we have 5,000 volts across both modulator and

amplifier, and the same current through each, but since they are in series the supply voltage is not E, as marked in Fig. 313, but 10,000, the point B. The power supplied is the same as before, the supply voltage being doubled and the current halved. Half this power is dissipated by the modulator and half delivered to the high frequency valve, the latter converting some 70% of its input into high-frequency output and dissipating some 30%. Thus the same anode-dissipation ratings as for the choke case will be necessary.

If now we consider the effect of an alteration of modulator grid voltage, this will sweep the modulator current in phase and the modulator voltage in antiphase along the load line and create amplifier anode-voltage and current changes, the modulator losing and the amplifier gaining power by an amount dependent upon the depth of modulation and type of signal. Thus the action is similar to the choke modulated case.

It is clear we shall have the same limits due to bottom bend and grid current and in consequence the maximum change of modulator voltage will be less than 5,000, and it is not possible therefore to modulate to 100%. In the case shown in Fig. 313, that is with equal voltage across modulator and high frequency valve in the quiescent condition, the modulation factor is about 80%. If, however, the initial quiescent condition is arranged such that the modulator takes greater voltage than the high-frequency valve, the modulation percentage can be raised to well over 90%.

There is no iron-core choke in the series circuit and the load line, therefore, is always substantially a straight line, and it is found possible to obtain modulation over a greater length of characteristic without introducing distortion. Measurements made with this type of control show that the distortion factor * can be kept within 3% even up to a modulation percentage of 90%, and in consequence the series modulation method is very suitable where high quality is required, or where the modulation involves a wide frequency spectrum.

* The distortion factor is defined as follows : If a pure sine modulating tone is applied to the apparatus and the output modulated wave analysed, then the percentage distortion factor is given by

$$\frac{\text{Root sum square of harmonic amplitudes}}{\text{fundamental amplitude}} \times 100.$$

Series Modulation, Cathode-Follower System [4, 5, 6]

We will now deal with the cathode-follower series modulation circuit, in which the modulator valve is at a H.F. potential to earth, one side of the H.F. circuit is earthed, and the modulating voltage applied between the grid of the modulator and earth and not between grid and cathode ; thus the modulating voltage is also across the load circuit. A simplified diagram of such a system is shown in Fig. 316, where V_1 is the series

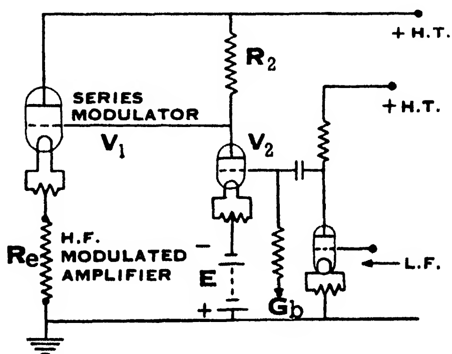


FIG. 316. Series Modulation (Cathode-follower).

modulator, and R_e represents the H.F. amplifier circuit whose equivalent resistance will be equal to a value

$$R_e = \frac{E_{av}}{I_{av}}.$$

The grid of the series modulator is fed from the anode of the valve V_2, in the anode circuit of which is a resistance R_2, the value of this being made appropriate to supply the correct voltage to the grid of V_1. If the series modulator stage consists of more than one valve in parallel, then the grid of each valve will be tapped separately along this resistance so that each valve shares the load equally. Between the cathode of V_2 and earth is included a D.C. supply with positive to earth so as to increase the H.T. potential on this valve, and so allow V_1 to modulate the main anode potential 100% without running into grid current limitations. The grid of V_2 is fed from an orthodox resistance-capacity amplifier between grid and earth.

Examination of Fig. 316 indicates that the currents in the valves V_1 and V_2 will be in opposite phase, for when the current in V_2 is cut off, the voltage applied to the grid of V_1 will be the least negative (relative to its cathode), whereas a rising current through V_2 will cause a fall of potential on the anode

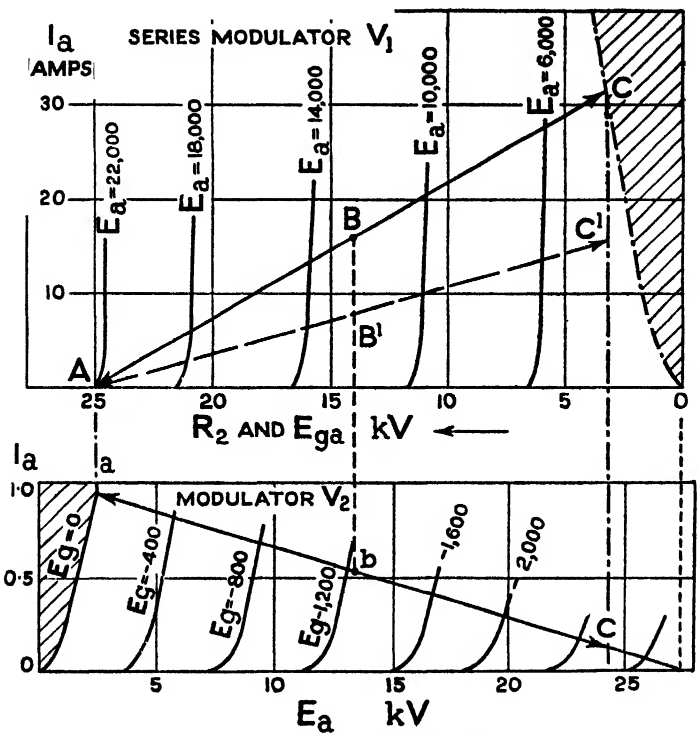


FIG. 317. Load Lines for Series Modulation.

of V_2 and in consequence an increasing negative on the grid of V_1.

In order to explain the working of the circuit clearly a definite case has been set out, the figures being taken from the Marconi 150 kW short-wave broadcast transmitter as supplied to the B.B.C.

In this case the carrier power to be modulated is 150 kW, and if the amplifier H.T. is to be 9,500 volts, the carrier current will be 15·7 amps. In order, therefore, to obtain 100%

modulation without distortion we have to arrange the series modulator so that we can obtain a peak current of 31·4 amps. without running into grid current. When the current is cut off, the modulator V_2 (which will then be passing full current) must not take grid current either. How this is accomplished can be seen by examining Fig. 317 where the characteristics of both valves V_1 and V_2 have been aligned one above the other so that the load lines can be directly correlated. From this figure it is observed that whereas the characteristics of V_2 are plotted in the usual way, namely as E_aI_a curves, those of V_1 have been plotted to show the relationship between the anode current and the voltage between anode and grid, instead of anode-cathode. The reason for plotting them in this way is that the voltage change between grid and anode is the same as that across the resistance R_2 and hence can be directly related to the curves of V_2; whereas because of the resistance load included in the cathode circuit and the fact that the grid potential is across grid and earth, the relation between I_a and E_a in V_1 is not so directly obvious.

Examination of these curves shows that the positioning of the curves of V_1 is such that the origin of voltage of V_1 is exactly above the current cut-off point of V_2, since at this point there is no voltage drop across R_2, and in consequence across E_{ga}. We cannot work down to this point, however, but only to such a point that no grid current flows in V_1 (the area shown shaded), indicated at *Cc*, this point being the peak of modulation. The limit at the other end indicated by *Aa* will be set by the $E_g = 0$ curve of the valve V_2, the excess voltage provided by E being sufficient so that the anode current of V_1 can vary from 31·4A to zero, without running into grid current on either valve. The carrier setting will be indicated by a point half-way between these limits, shown at *Bb*.

In order to indicate the actual voltages across the various parts of the circuit for the peak, carrier, and trough conditions, figures have been extracted from these characteristics and set out in diagrammatic form in Figs. 318 and 319. In Fig. 318 the figures placed between arrows indicate difference of potentials, and figures by themselves indicate potentials above earth. It can be seen that the potentials on the modulator V_2 exceed those on the series modulator V_1, although of course there is a

power gain from the V_2 stage to that of V_1. Fig. 319 sets out the variation of voltage and current with time for the same peak conditions, it being assumed that a sinusoidal voltage is

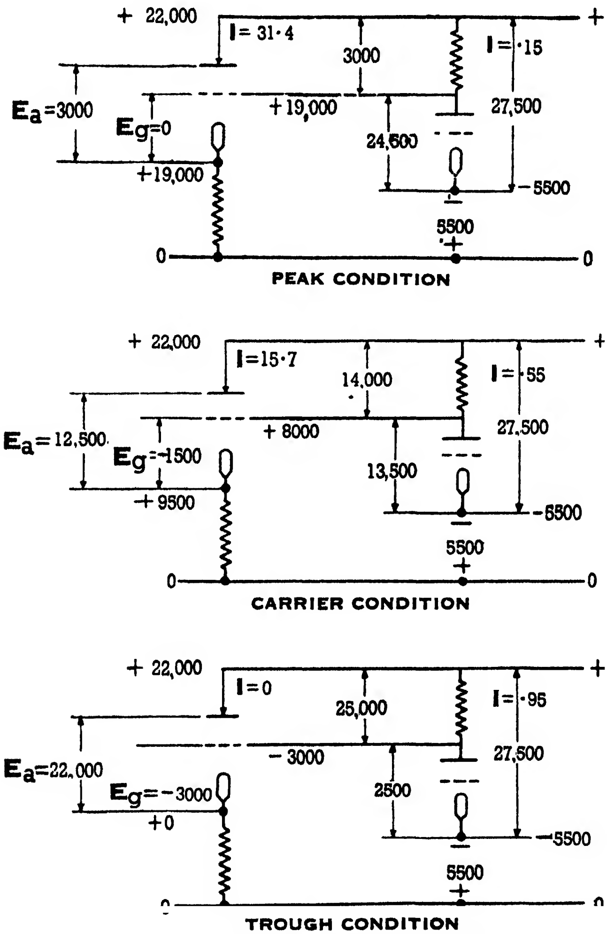


FIG. 318. Illustrating Series Modulation.

applied to the grid of the modulator. These curves show clearly that the anode of the series modulator remains at a constant potential above earth and that the other potentials vary relatively to it, and to earth. The voltage changes, grid-cathode, grid-earth, cathode-earth, all vary in phase with one

another and with the change of I_a. Of course, from the conversion point of view this phase relationship is the same as that of an ordinary circuit, as the anode voltage and anode current are virtually in anti-phase, if we consider the cathode as a datum. It will be noticed that the changes of cathode potential are the same as those of the grid potential (relative to earth) both in phase and amplitude (nearly), hence the name cathode-follower. They would be the same in amplitude if a very high-μ valve was used.

The inclusion of the load resistance in the grid circuit means that there is no voltage magnification, but on the contrary a

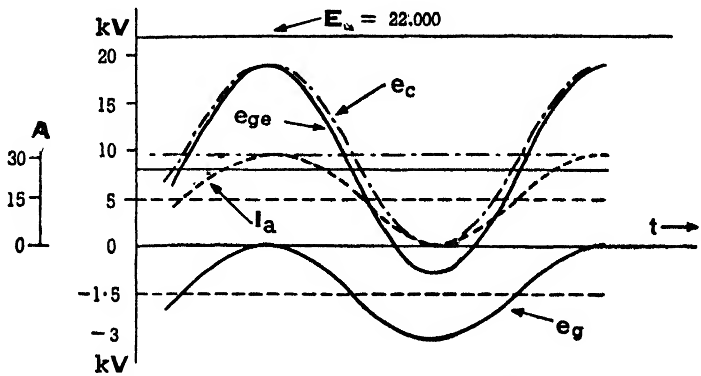


FIG. 319. Current and Voltage in Series Modulation.

small reduction ; it is, therefore, equivalent to a 100% negative-feed-back circuit, and as such it has a number of advantages where high linearity is desired, amongst the advantages being the following :

(1) With a linear grid-swing, non-linearity of the series modulating valves will have a negligible effect.

(2) With a non-linear load the voltage swings are still linear.

(3) A variation of load resistance over wide limits will not affect the modulator setting appreciably. This can be seen by considering Fig. 317 where the load line for a load of double the resistance value, shown at $AB'C'$ has been added.

Other advantages are that A.C. lighting for the series modulator can be used without any hum ripple appearing, variations due to H.T. changes are negligible, and with paralleled modulators the removal of one valve will not alter the operation of the circuit. Actually it is probable that the future of such circuits will be more for low-power, high-linearity circuits than for high power.

The foregoing discussions have been confined to the modulation of any single stage by anode modulation of that stage, it being assumed that the grid of the H.F. stage was supplied from a constant amplitude H.F. source.

Low Power Modulation

As mentioned previously modulation may be carried out at the final stage of an amplifier, in which case each H.F. stage has a constant input voltage. If, however, modulation is effected at a low level, all subsequent stages have to handle a modulated carrier and thus require a different setting. In Chapter X it was shown that the input-output characteristic of any Class B amplifier can be made to have a linear relationship between input volts and output H.F. current, and the curves of Fig. 217 were given as an example. This relationship of I_{hf} and input is shown again in Figs. 320, 321 which also indicate the change of efficiency of the amplifier stage.

Thus if we set such an amplifier to the point A statically (Figs. 320, 321), thus making it a Class B amplifier, and drive it up to the point O by high frequency impressed on the grid (representing the carrier condition), any modulation impressed on the input high frequency will lead to modulated high frequency in the output, and if the input frequency is modulated between limits of OA, OB, we shall get a high-frequency output 100% modulated.

This modulation can be carried out in one of two ways: by keeping the bias constant at A and varying the amplitude of input as shown in Fig. 320; or by keeping the input H.F. amplitude constant and varying the bias by modulation about the point A as shown in Fig. 321. In either case the result is the same, modulated H.F. output linearly related to modulated input. Observe in both cases that when driving to the point O, i.e. the carrier, or unmodulated condition of the amplifier, the

efficiency is but half that obtained under full or peak driving conditions. Hence modulation is achieved by a variation of amplifier efficiency. Thus the average efficiency of such a

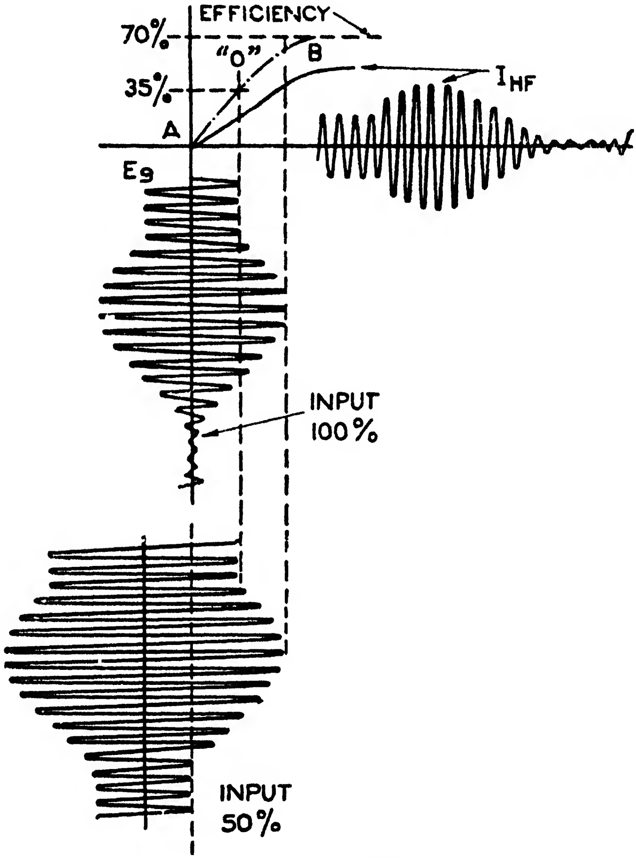


FIG. 320. Operation of Modulated Class B Amplifier.

modulated stage is but half that of a stage driven at constant input voltage up to peak values.

This is shown in Fig. 322, where the E_aI_a curves of the amplifier are shown, and the cyclic change of voltage, current and efficiency, for sine modulation, Fig. 320 showing the connection between the input modulated wave and output.

A second point to observe is that the depth of modulation

of one stage has no relation to that of the other. The input may be deeply modulated, and yet no modulation appear in the

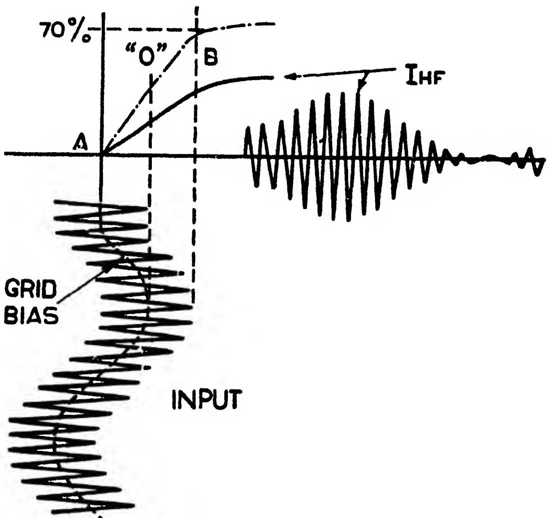


FIG. 321. Grid Modulation of Amplifier.

output. This would occur, for instance, if the stage was set too high in the carrier condition, as *B*, above.

On the other hand, it is possible to obtain a 100% modulation

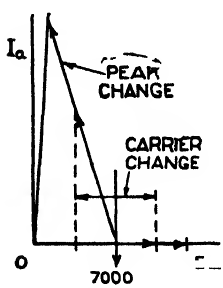


FIG. 322a. Load Line for Modulated Amplifier.

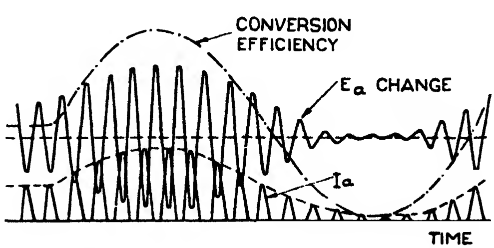


FIG. 322b. Current and Voltage in Modulated Amplifier.

of output even if the input has but shallow modulation. For instance if the amplifier is biassed very negative, say beyond the point A as shown, thus making it a Class C amplifier, and the input H.F. is increased to such an amplitude that it still drives the amplifier up to the same point O in the

carrier condition, then only moderate control of this larger input (50% in the case shown) will still give full modulation. Such a condition is, however, not usual owing to the difficulty of controlling the input to within the prescribed limits.

As has been mentioned, we can obtain the same result by keeping the input excitation H.F. voltage constant and varying the bias at the modulation rate as shown in Fig. 321. The former method has already been described, and the second method which we will now discuss can be carried out in one of two ways.

Grid-Current Modulation

In this arrangement the grid-bias changes are produced by introducing a valve as a grid resistance. By varying the resistance of this " valve " leak at the modulation rate, the bias on the stage to be modulated will be changed. It is not a simple matter to obtain linearity by this method, but it is useful for certain classes of work and it has not the frequency limitation of the choke methods.

Grid-Voltage Modulation

D.C. grid voltage control is but slightly different. Instead of changing the grid bias by varying the resistance value in the grid circuit, the modulation is impressed across a grid circuit resistance, the voltage drop across the resistance due to modulation controlling the grid bias at the modulation frequency.

Included in the D.C. grid circuits of the amplifier valves is a resistance which is of low value, such that the normal grid current of the two valves produces a negligible negative static bias. Across the resistance is placed the modulator valve and its anode supply voltage, both carefully insulated from earth. It will be clear that if the anode current of the modulator is large compared with the grid current of the amplifier, the voltage drop produced will be a function of the modulator valve feed and in consequence the negative bias produced on the H.F. valves will follow directly changes of modulator feed.

Thus if the modulator is set to the centre of its characteristic in the quiescent condition, application of the signal voltage to the modulator will create similar changes of amplifier grid negative bias. Provided, therefore, the initial bias is such as

to set the H.F. amplifier to the centre point of its output curve as previously explained, linear change of H.F. output will result.

This system has been very successful in dealing with high modulation frequencies and has been used for a television transmitter having a frequency response up to two megacycles.

Cathode Modulation [6]

A popular system of modulation for small power sets is the so-called "Cathode-Modulation" system, where the modulating E.M.F. is injected between the cathode and earth, the grid and anode circuits being returned to the earth side of the modulation system. The circuit of a cathode-modulated stage is shown in Fig. 323, where included in a conventional driven stage is a modulation transformer M with its secondary circuit connected between the cathode of the amplifier valve and earth, the primary being supplied from a normal push-pull modulator.

Assuming the H.F. stage is being driven from a constant-amplitude source in the usual way, then the application of an A.C. modulation potential across the secondary of M will produce a voltage which will vary the cathode potential about earth. Or, if we consider the cathode as a reference point, then the common point of connection of grid and anode (namely E) will vary at low frequency as regards the cathode, so that as the anode voltage increases above the D.C. value, relative to the cathode the grid will become less negative, and as the anode potential falls the grid potential will go more negative. Thus such a type of control is a combination of both grid and anode modulation, the amount of control being determined by the voltage supplied from the modulator and the value of grid bias resistance, R_g, Fig. 323, provided. That is to say, we can increase the proportion of anode modulation relative to grid modulation by increasing the modulation voltage (and the size of the modulator); and making R_g, and in consequence negative bias, large. By reducing the grid bias a smaller applied modulating-voltage will have a greater control of the grid potential and the system becomes nearer to a low-level type.

Claims have been made that it is possible to obtain a conversion efficiency as high as 50% with 100% modulation, and

although it appears fairly easy to set up a circuit to give such a figure, examination of the wave form will indicate a considerable harmonic content. Some experimental figures obtained at the School of Wireless Communication at Chelmsford indicated that to obtain a circuit with less than 2% harmonic content, it was not possible to get a conversion efficiency greater than 35% for 100% modulation, although, of course, if

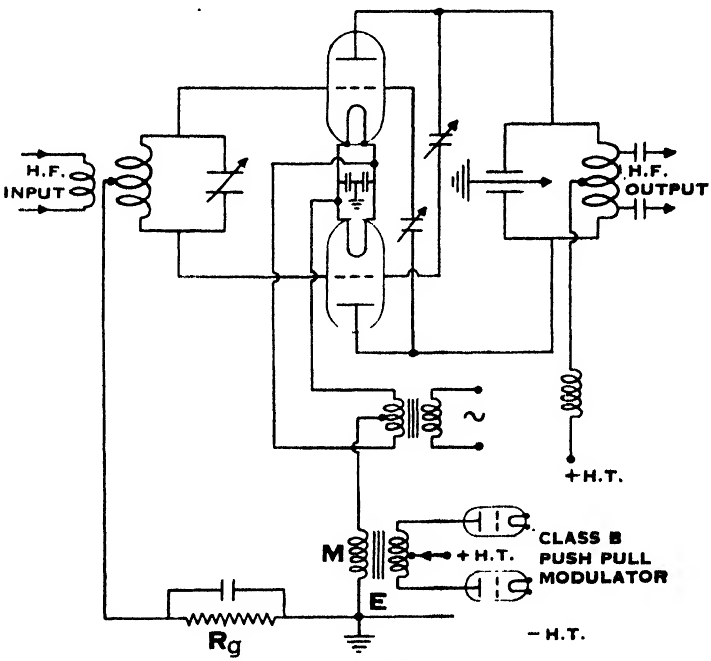


FIG. 323. Cathode Modulation.

the modulation percentage is reduced, and the circuit readjusted, the conversion efficiency can be increased. Fig. 324 shows that a very linear relationship between modulator input and percentage modulation can be obtained, and it is found that under the maximum percentage-modulation condition the modulation input power is but 10% of the output power. A small point to be observed is that the driving voltage needs to be about 50% of that necessary to drive the same amplifier stage as a Class C amplifier to full output, although the grid

bias required is greater, being some six times the cut-off bias, as against the more normal value of 3·5. This means we are virtually operating the amplifier as a much under-driven, Class C amplifier. The frequency response of the

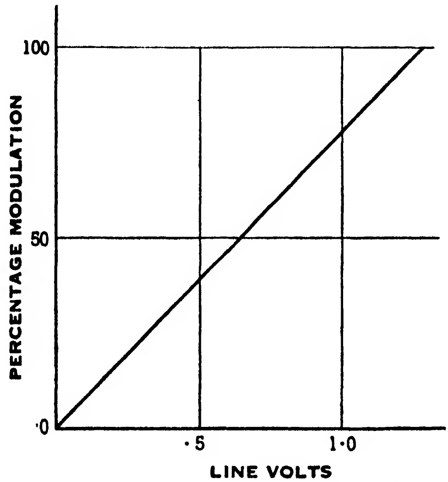


FIG. 324. Input-Output Curves for Cathode Modulation.

circuit is good, being within 1 db over a frequency range from 100 to 10,000 cycles, without any special precautions being taken, although it is found the load coupling is fairly critical.

Suppressor-Grid Modulation for Pentode Amplifier

On page 403 it was mentioned that one of the advantages of the pentode amplifying valve over the triode was the ability to use a simple system of low-power modulation. This is carried out by the application of the modulation to the suppressor grid. Fig. 325 shows the relationship between suppressor-grid volts E_{g3} and output current $I_{H.F.}$ from which it is seen that with the suppressor grid somewhat positive, full output is obtained at a high conversion efficiency, and as the suppressor grid is made negative both output current and efficiency fall, there being a linear relationship between E_{g3} and $I_{H.F.}$ over a considerable portion of the curve.

Thus if we adjust the value of E_{g3} to say $-$ 65 volts for the carrier condition, we can sweep to zero volts for the peak of

modulation and to − 130 volts for trough, and obtain a linear relationship throughout. This can be accomplished by in-

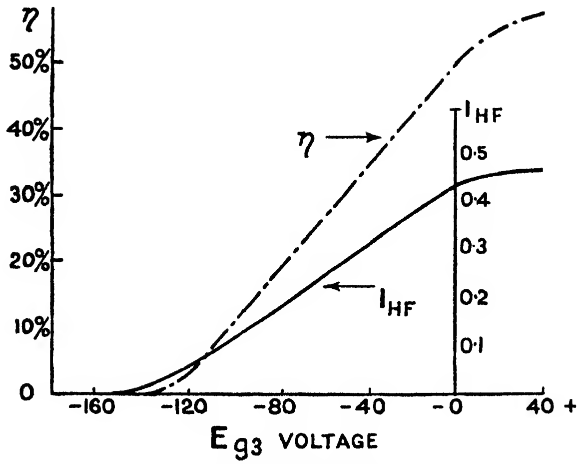


FIG. 325. Output/Suppressor-grid Voltage Curve.

cluding in the suppressor grid circuit, see Fig. 328, page 539, a modulation transformer and the necessary bias supply in series.

The Keying of Telegraph Transmitters

The keying of a telegraph transmitter is, of course, a special modulation problem, but since only a clear distinction between the " mark " and " space " condition is required, we do not need to use " high fidelity " modulators for such a purpose but simpler circuits are possible. Although the square wave represents the ideal telegraph signal since it has the greatest ratio of mark to space, it also creates an unnecessarily wide frequency spectrum. Thus a rounded shape of mark, usually determined by the circuit constants and signalling speed, is to be preferred for this reason even though the mark/space ratio is smaller and the effective signal therefore less. Although the waveshape may be predetermined by the proper design of shaping circuits, with most transmitters it is somewhat fortuitous. The *C*, *R* and *L* constants in the circuits and the non-linearity of the valve characteristics result generally in a waveshape of exponential form.

Apart from the waveshape produced, the principal require-

ments for the satisfactory keying of a radio transmitter may be summarised as follows : No change of frequency due to keying, negligible radiation during space, no large voltages surges, no interference in neighbouring receivers tuned to a different frequency, and the radiation of no transient " noises." It is not easy to satisfy all these conditions, particularly by keying at a single point. Thus the nearer we key to the master oscillator the less the space wave but the greater the tendency to frequency changes. Except for the very simplest of transmitters, keying is usually operative at more than one point.

Voltage surges will occur if the keying arrangements are such that the power drawn from the supply is very different during mark and space and if the power supply system cannot follow these changes with sufficient rapidity without surging.

Interfering noises, termed " key clicks," may be heard in neighbouring receivers even though these are tuned to quite a different frequency from that of the transmitter. This is because the starting and stopping of radiation at the beginning and end of a mark is too sudden, so that the transient contains a great number of widespread components. Where the keying contacts have to control a large voltage it may be difficult to avoid some degree of sparking, causing damped oscillations in the circuits affected and the radiation of " noise." The key-click trouble is only serious from high-power transmitters and may necessitate the control of keying waveshape. This must be carried out near the output stage since any control of wave-shape in the earlier stages may be ineffectual due to succeeding circuit constants.

For the keying of low-power transmitters at hand speed very simple keying arrangements are satisfactory, as, for example, a key in the supply system, but with most transmitters where any speed of signalling is required or any large power is handled, more elaborate methods become necessary, and control has usually to be effected at more than one point as mentioned previously.

It will be appreciated that the cutting " off and on " of a transmitter abruptly, tests the regulation of the supply severely and with high voltages, transient effects may be induced which impose a severe strain on circuit components. In such cases it will be necessary to provide an alternative load during space,

known as an absorber circuit, the amount of back-load required being largely dependent upon the type of supply used. Thus if a hard-valve rectifying system is employed, for example, then the regulation will be poor, that is, there are large changes in the D.C. supply voltage when the load changes. Also, due to the large smoothing system, the " electrical inertia " of the arrangement is large and cannot respond quickly to sudden changes of load.

The regulation of hot-cathode mercury vapour and mercury arc rectifiers is, however, much better, and so is that of the large D.C. generator. With such, therefore, only partial absorption will be necessary, the actual amount being dependent upon the actual circuit and keying requirements. With the larger transmitters it is usual to employ a separate absorber circuit, the back-load energy being dissipated partly in a resistance and partly in the absorber valve itself, but with some of the smaller transmitters it is possible to employ the amplifying valves themselves as partial absorbers during the spacing periods.

We will now describe a few typical circuits showing different methods of keying.

The Partial-Absorber Method of Keying

There are many varieties of this circuit, one of which is shown in simplified form in Fig. 326, where the H.F. circuits

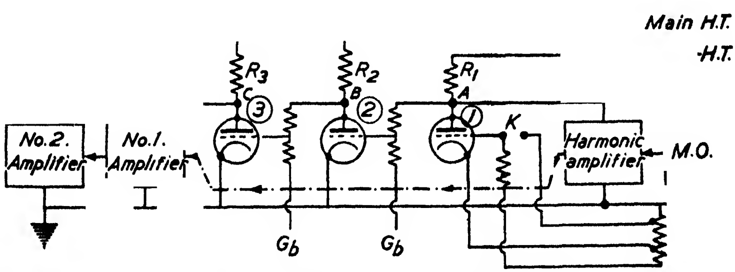


FIG. 326. Partial-absorber Keyed Circuit.

are merely indicated in block form, and the keying circuits indicated at 1, 2 and 3. With this particular transmitter the master oscillator and harmonic amplifier are run from a separate power supply, to minimise frequency changes, and the keying circuit is arranged to operate on both. Thus

the keying valve 1 controls the voltage to the harmonic amplifier and at the same time controls the main H.T. to the first power amplifier, by control of the partial absorption circuit 3 on space. Since the power handled by this absorption circuit may be fairly large, a sub-absorber, 2, has to be inserted between the keying valve and main absorber.

Control is accomplished by a normal polarised relay (the contacts of which are indicated by K) operated by the incoming tone or D.C. from a line or locally from a hand key, *K* being open for "mark" and closed for "space." It will be seen from Fig. 326 that the cathode of the keying valve 1 is connected to a point on the bias circuit negative of earth, but the grid can be made either more negative or less negative than the cathode through the alternative paths indicated. Thus with the contacts of *K* open (mark) the keying valve grid will be more negative than its cathode, but with the contacts closed the grid will be positive relatively. In the former case no

TABLE XXII

Key.	Mark.	Space.
	Open.	Closed.
Keying Valve (1) .	No valve current. Volts at *A* rise. H.A. fully operative, driving No. 1 amplifier.	Valve current. Volts at *A* fall. H.A. output reduced, very small drive to No. 1 amplifier.
Sub-absorber (2) .	Grid positive, due to high voltage at *A*. Valve current. Fall of volts at *B*.	Grid negative, due to bias. No valve current. Rise of volts at *B*.
Absorber (3) . .	Grid negative, due to bias. No absorber current. Rise of voltage at *C*.	Grid positive, due to rising voltage at *B*. Absorber current, power dissipated in *R* and valve (3). Fall of voltage at *C*.
No. 1 amplifier .	Full volts on anode due to rising voltage at *C*. Grid driven by full output of H.A. Full drive to No. 2 amp.	Reduced voltage on anode and lack of drive from H.A., reduces output to negligible value.

current will flow in 1, but in the latter case (space) the keying valve takes current. The effect of this current change in valve 1 on other parts of the keying circuit can best be shown in tabulated form (page 536).

With such an absorber system, a transmitter of the highest power can be keyed at high speeds from a light relay operated by incoming line singals.

As an example of a circuit in which the amplifying valves act as absorbers. Fig. 327 is typical. From this it is observed that there is a resistance between cathode and the negative H.T. of the last two amplifying valve stages (marked " No. 2 mag." and " No. 1 mag."). The keying relay leaves this

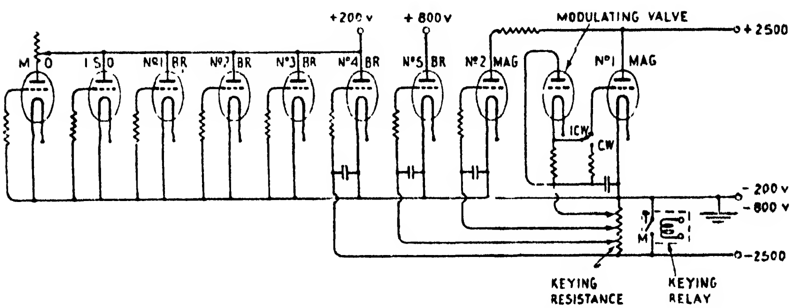


FIG. 327. Absorber Keying using Amplifier Valves.

resistance in circuit during space but shorts it during mark. During space, therefore, the feed of these two amplifiers flows through the resistance, thereby providing self-bias but also providing sufficient grid bias to the stages marked " No. 4 *BR* " and " No. 5 *BR* " to cut them off. Hence the last two stages are not being driven, there is no output from the transmitter, and these stages take a static feed of value depending upon the tapping on the keying resistance.

Since the feed of any given valve operating under Class C conditions can be considerably greater than its static feed it is clear that a keying system of the type described can only absorb a small proportion of the total power. And since only the last valve or two are of any size, it is no value utilising the other valves in the chain for the purpose. This means that unless the regulation of the supply is reasonably good, voltage

fluctuations will occur in the set causing possible breakdowns and frequency scintillation.

For the smaller transmitter where valve absorber keying is not necessary, many possible methods are available, and have from time to time been adopted. The amplifier, or oscillator, or both can be keyed, although with a crystal driven circuit it is preferable not to key the oscillator. Since keying has to make any stage "dead" or "live," control can be effected by a simple change of potential in any one circuit, or more. Thus either anode or screen volts can be reduced or removed or a change made in the static potential of a control grid. Naturally enough such changes can be made by the use of either shunt or series control units and either valves or relays employed. As a general rule, anode keying is no longer employed since the voltages to be handled are the maximum and an alteration of screen or control grid potentials is now more usual.

In small installations, which usually operate "simplex," the keying circuits of the transmitter often incorporate auxiliary circuits, so that "break-in" or "listening-through" can be adopted. During each space the receiver is fully sensitive and connected to the aerial, but during each mark the receiver is protected from the transmitter voltages. Thus the station being communicated with can gain the attention of the operator if necessary.

When a single aerial is being employed for both transmission and reception, the solenoid-operated send/receive switch acts as a keying relay and its groups of contacts carry out the required circuit changes in the correct sequence. Thus in passing from mark to space the relay, which would have four pairs of contacts, will :

(1) Key transmitter from mark to space.
(2) Change aerial from transmitter to receiver.
(3) Remove short on receiver input (or output).
(4) Change telephone from side-tone circuit to receiver.

The provision of an artificial side-tone circuit is necessary when the receiver is completely de-sensitised and will usually be provided by connecting the phone circuit to the modulator of the transmitter.

When two aerials, one for transmission and one for reception,

are used, the circuit changes can be greatly simplified, as during the mark periods it is only necessary to reduce the sensitivity of the receiver to a low level, say by the application of negative bias to various control grids. By leaving a small output from the receiver during the mark periods, the required amount of side-tone can be obtained. Since the change-over from mark to space can be affected by a simple change of bias, key clicks can be entirely avoided, whereas with the keying relay system the avoidance of key clicks is dependent upon the adjustment of the contact sequence set up.

We will now show a few examples of the keying circuits of some small transmitters employing different methods.

Typical Keying Circuits for Small Transmitters

Fig. 328 shows the basic keying circuit of a transmitter for about 100 watts employing a triode master-oscillator and

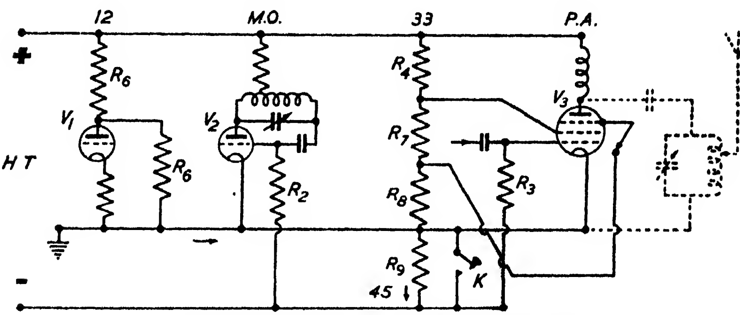


FIG. 328. Potentiometer Keying Circuit.

pentode amplifier, with a triode modulator which can be used for telephony or made to self-oscillate to provide a tone for M.C.W., and for the keying side-tone where "break-in" is provided. It will be observed that across the power supply is a potentiometer made up of resistances R_4, R_7, R_8 and R_9, earth being at the junction of R_8 and R_9, the latter resistance being about 5,000Ω. The M.O. and P.A. grids are joined to the negative end of this resistance as shown, a pair of contacts on the keying relay being connected across R_9.

In the space condition, with the key open, there is a total feed through R_9 of about 45 mA, from the potentiometer and modulator, thus making the control grids of both M.O. and

P.A. some 230 volts negative to their cathodes. This is sufficient to cut off the feed to both valves, thus stopping the

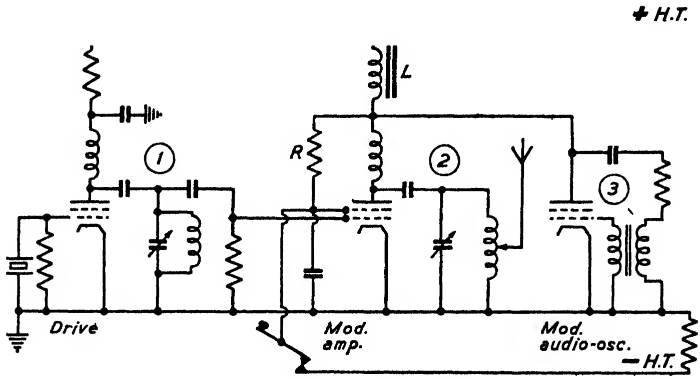


FIG. 329. Amplifier Keying.

drive oscillating and the P.A. from taking feed. In the mark condition, since R_9 is cut out of circuit by the closing of the

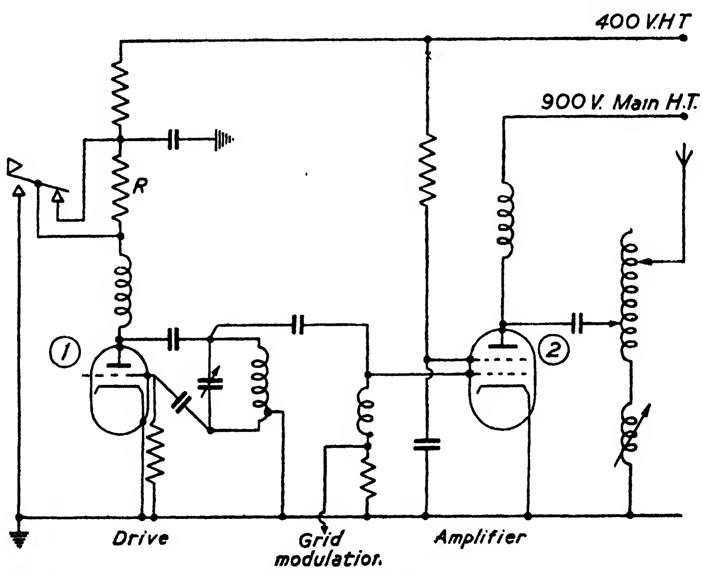


FIG. 330. Drive Keying.

keying contacts, these grids are returned to earth, the valves merely take self-bias voltage due to their grid currents, the

M.O. now oscillating and the P.A. taking a feed which is dependent upon the setting of the suppressor grid and screen volts, determined by the potentiometer design.

Fig. 329 shows a small transmitter designed for phone, *MCW* and *CW*, employing screen-grid valves throughout. The drive is crystal controlled and remains operative all the time, the key merely biasing negatively the screen of the amplifier valve (2). It will be observed that there is a D.C. current through the speech choke L and resistance R at all times which remains substantially constant on both mark and space, and this eliminates surges.

Fig. 330 shows the circuit for a 15-watt transmitter employing a triode self-oscillator drive (1) and screen-grid amplifier (2) in which only the drive is keyed. In the space condition the drive anode is earthed through the additional resistance R to maintain constant load on the H.T. supply.

Circuits for Frequency and Phase Modulation

To produce frequency modulation we need to vary the carrier frequency f_c by an amount Δf_c (the deviation) which depends upon the amplitude of the modulating signal, and to do this at the modulation frequency f_m. It will, therefore, be necessary to apply F–M to the master oscillator but at the same time we must ensure that the *mean* frequency is kept very constant.

Consider an oscillator in which the frequency f_o is determined by an inductance L and capacitance C. Then if the modulation can be made to vary either L or C in accordance with its amplitude and at its frequency f_m, a F–M output will be obtained. The reactance valve, to be discussed in the next section, is the usual way of achieving the required variation.

Since F–M is generally used in the U.S.W. band, the oscillator frequency f_o will usually be much lower than the final carrier frequency f_c. Suppose, for example, that f_o is 2·5 Mc/s, with a deviation of 3·75 kc/s at 100% modulation and that f_o is multiplied up 20 times to give an f_c of 50 Mc/s. Then at the instant when the modulation has made the oscillator frequency 2,500 + 3·75 kc/s, the radiated frequency would be 20 × 2,503·75 or 50,075 kc/s. Hence the deviation is 75 kc/s, or 20 times that applied to the oscillator. These figures are

typical of those which might obtain in high-fidelity F–M broadcasting where, as seen in Chapter III, it is desirable that Δf_0 should be several times greater than the maximum f_m.

Phase modulation can be produced if either the inductance or capacitance in the resonant circuit of an amplifier being driven from a constant frequency M.O. at the resonant frequency of the amplifier-timed circuit are varied, the reactance valve again being employed. Hence P–M has the advantage that it is not applied to the oscillator and there is not the same difficulty in keeping the mean carrier frequency constant.

We saw in Chapter III that F–M and P–M are closely related and that for sinusoidal modulation $\phi = \Delta f/f_m$. Consequently,

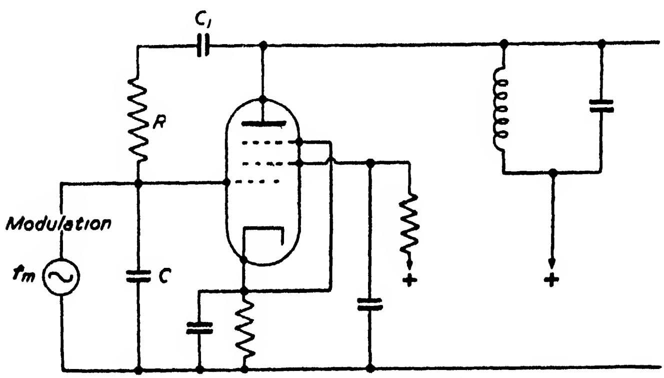


FIG. 331. Reactance Valve.

if the phase change made for all values of f_m is the same, then the deviation frequency produced will be directly proportional to f_m. If we wish to radiate a wave which is frequency-modulated with a deviation the same for 100% modulation of any value of f_m, then we can produce this result by phase modulating one of the amplifying circuits but in such a way that the phase change ϕ is proportional to $1/f_m$.

Taking the same figures as before, if P–M is applied to a stage working at 2·5 Mc/s and the final result is again to be "level" F–M at 50 Mc/s, with a deviation of 75 kc/s, the phase deviation for an f_m of 300 kc/s will be $3{,}750/300 = 12{\cdot}5$ radians and for an f_m of 3,000 would be 1·25 radians.

It is not found that deviations exceeding about one radian

can be obtained without distortion and in our example it would be necessary to increase the frequency multiplication.

If "pre-emphasis" is to be employed, that is, the higher values of f_m are to produce a larger deviation, then this reduces the values of ϕ for the lower frequencies and simplifies the provision of P–M.

Reactance Valve

This valve provides a reactance across the resonant circuit, which varies with the amplitude of the modulating frequency applied to the grid. Hence, if the resonant circuit is in an oscillator, frequency modulation results, whilst if it is an amplifier, phase modulation is produced.

Fig. 331 shows a reactance valve connected across a tuned circuit in its anode, modulation being applied to the control grid of the reactance valve.

The condenser C_1 is large and serves merely to block the D.C. anode voltage from the grid. R and C, however, form a phase-shifting combination, so that a proportion of the R.F. voltage across the resonant circuit is applied to the grid, but in quadrature with the anode voltage.

If we assume that the alternating current taken by the valve is $g_m E_g$ and that this is large compared with the current through CR, then the impedance, Z, which the valve places across the resonant circuit is given by $E/g_m E_g$.

But $$E_g = E\,\frac{-j/\omega C}{R - j/\omega C}$$

and hence

$$Z = \frac{R - j/\omega C}{-jg_m/\omega C} = \frac{\omega CR - j}{-jg_m}$$

$$\simeq \frac{j\omega CR}{g_m}$$

Hence the valve behaves as an inductance of value CR/g_m.

If now g_m is made to vary by the modulation voltage, it will be seen that the tuning of the resonant circuit will vary accordingly.

Examples are given in Chapter XVIII of reactance valves employed to produce frequency and phase modulation.

The Frequency and Phase Modulation of Telegraph Transmitters

Under certain ionosphere conditions it may be advantageous to spread the signal band by the superposition of modulation other than that caused by the telegraphic mark and space. Formerly this was carried out by amplitude modulation, either sinusoidal in character, or of more complexity. Such a method, however, has the disadvantage that the average signal strength is reduced as the tone modulation is increased. We can, however, apply frequency or phase modulation which will give any required spread of side-bands, but since the transmitter can now work at peak conditions throughout, the transmitter efficiency is not lowered as it is with tone modulation. Where the modulator is applied direct to an oscillator, a frequency modulation will be effected, but as certain types of master oscillator circuits do not lend themselves to frequency modulation of the order required for anti-fading work, it may be more convenient to employ a phase modulator. Whichever is used, the modulation index or phase angle should be such that there is a fairly uniform distribution of energy between the carrier and first four side-bands. This is obtained with a modulation index of 2, and as a modulation tone of the order of 400 c/s is commonly adopted, the frequency deviation at the carrier frequency will be 800 c/s. Most telegraph transmitters operating in the short waveband, i.e. between 3 and 22 Mc/s, usually employ a M.O. of about 1 Mc/s whether of the *LC* circuit or crystal type, so that at the lowest radiated frequency since the frequency multiplication is 3·0 the phase angle will be 2/3 at the M.O., i.e. ·66 radian, a workable figure for phase modulation.

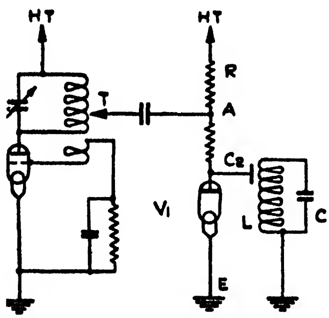


FIG. 332. Frequency Modulation of Franklin M.O.

The Franklin drive design lends itself directly to frequency modulation and one method of applying this is shown in Fig. 332, where *LC* indicates the oscillatory circuit of the Franklin drive unit. It will be seen that a plate C_2 is coupled to the inductance giving a minute capacity coupling. This

plate is connected to earth through a diode valve V_1 and it is clear that any variation of feed through the diode will vary the earth capacity effect of C_2 to the inductance, and this in turn will vary the frequency of the master oscillator. Fig. 333 gives an idea of the relationship between diode feed and frequency change, from which it is clear that if we wish equal changes in anode current to produce equal changes in frequency we must set the diode on the straight part of the curve (O, say), by supplying a D.C. potential. The steepness and straightness of the curve connecting frequency shift and diode current depends largely upon the position of C_2 along the inductance.

The source of the modulation is a note oscillator of frequency f_m the output from which is applied to the diode circuit (see Fig. 332) so that the diode current is varying at a frequency of f_m about the mean value. Hence the frequency of the master oscillator is made to vary between certain limits $f_c + f_n$ and $f_c - f_n$ the change occurring f_m times a second.

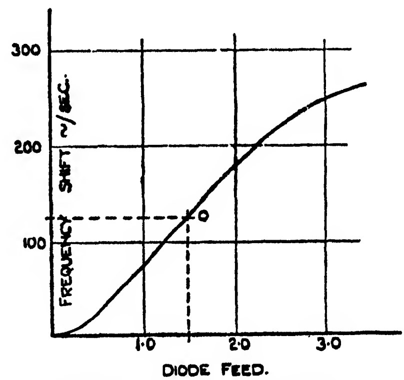


Fig. 333. Frequency Deviation of Diode-controlled M.O.

The frequency change f_n will evidently depend upon the amplitude of the voltage applied from the note oscillator since this determines the amount of variation of diode current and the position of the tap T will, therefore, control the value of f_n.

The changes indicated above are at the M.O. frequency, but the actual deviation frequency on the carrier will be greater by the frequency multiplication used.

A telegraph transmitter employing frequency modulation will be keyed in the normal way in a stage subsequent to the master oscillator (usually by the absorber method) so that the transmitter is still working on the principle " full radiation on mark "—" no radiation on space," but the radiation on " mark " takes place on a small band of frequencies.

Phase-Modulated Telegraph Transmitter

Such a method is not applicable to a crystal type M.O. and for this type of drive a phase-modulated circuit is preferable. This can be accomplished by the introduction, into an amplifier stage, of a reactance valve on whose grid is applied a tone signal, the circuit being shown in Fig. 334, where *LC* is the tuned circuit of an amplifier stage driven from the crystal oscillator at a frequency of f_o, *V* the reactance valve with a 400-cycle tone applied to its grid, the frequency-modulated output being taken from the anode circuit. In order that the valve shall behave as a reactance the voltages on grid and

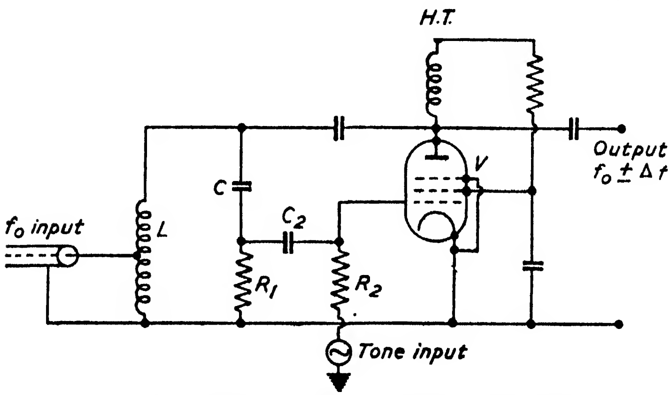


FIG. 334. Phase-modulated Telegraph Transmitter.

anode must not be opposite in phase, but have a quadrature component, and this is accomplished as shown with the example discussed previously in the chapter, by the circuit C_2, R_2 across R_1. It is claimed that for this type of work a phase modulator is superior to the frequency modulator, as with the latter mains hum is likely to produce considerable unwanted modulation since with frequency modulation the lower the modulation frequency the greater the phase angle or modulation index; whereas with phase modulation the index is a constant.

Circuits for Pulse Modulation

The principle of pulse modulation has been discussed in Chapter III. This method of communication is sufficiently new that practice has not become standardised. Space does

not permit of treating even briefly the various methods which have been employed for producing pulses which are amplitude, duration, phase or frequency-modulated. We shall, therefore, deal with only one of the simplest circuits for producing pulse-duration modulation.

The valve of Fig. 335 has a short grid base and the sine wave input (which is at the desired recurrence frequency of the

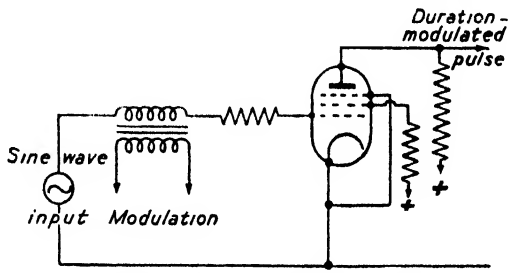


FIG. 335. Pulse-modulation Circuit.

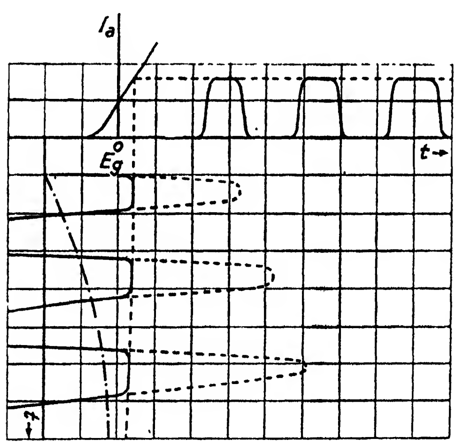


FIG. 336. Pulse-duration Modulation.

pulses) is large, so that the grid swings beyond cut-off and also goes positive.

The amount by which the grid itself goes positive is nearly constant, irrespective of the magntiude of the sine wave input, because grid current drops the voltage in the series resistance.

In the absence of modulation, therefore, the anode current would consist of a series of nearly rectangular pulses.

If a sinusoidal modulation is put in through the transformer then the grid voltage and resulting anode current are shown in Fig. 336. It will be seen that the duration of the pulses of anode current now vary in sympathy with the modulation.

CHAPTER XIV

RECEPTION OF SHORT AND ULTRA-SHORT WAVES

In this chapter it will be assumed that the reader is familiar with the ordinary principles of reception, and has a knowledge of valves, valve amplifiers, frequency-changers and detectors such as may be gained from one of the more elementary text-books on radio. We propose here to give a preliminary survey of the various problems, and describe the outline features of a few basic types of receivers such as "Straight" "Super-regenerative," "Super-heterodyne," stressing the features which are important in short and ultra-short wave work. Special receivers used for frequency modulation and centimetre-wave reception will also be described. In addition there are the special problems associated with radar receivers.

The various classes of traffic (as indicated below), which may have to be handled by a receiver, show the variety and complexity of the general problems of design.

(*a*) Broadcasting.
(*b*) General purpose telegraph and telephone reception by a skilled operator.
(*c*) High speed telegraph signals for relaying and recording at a Central Telegraph Office.
(*d*) Telephone signals suitable for linking together the public telephone networks of various countries.
(*e*) Television.

Generally speaking receivers evolved for duty in any one of the above-mentioned groups will only be suitable for that particular purpose and it is not an economic proposition to develop a universal type of receiver.

Those used for (*b*) are usually referred to as communication receivers and those for (*c*) and (*d*) as commercial receivers, though in recent years there has been a tendency (probably

temporary) to use the higher grade communication receiver for these purposes.

In all receivers the following are desirable features: high output signal/noise ratio, selectivity, fidelity, sensitivity and ease of control.

How near the receiver approaches the desired ideal will depend upon the purpose for which the receiver is to be used and upon any economic considerations which may be imposed. For instance receivers in (*a*) and (*e*) are for entertainment purposes and must give pleasing reproduction, have a very simple system of control, are often required to work with any aerial, but they need have only low sensitivity as they will work from high field strength levels and probably in noisy localities. Receivers in (*b*) will have to be capable of quick searching, and even in the present state of the art will often have to work with simple transmitters which are not too constant in frequency. Sensitivity may have to be high or low and as both telegraphy and telephony may be used, methods of controlling selectivity are desirable, and also a variable time-constant for the automatic volume control is necessary. Evidently the requirements of (*c*) and (*d*) are very exacting and the nature of the services so important that elaborations are here admissible which would be out of place in the other groups.

The peculiar problems of (*e*) are in providing the very large band-width necessary and in keeping down phase distortion.

Noise

Noise may be produced by causes which are external to the receiver or internal. Atmospheric interference with reception has already been discussed (page 158) and it has been pointed out that interference is caused on all frequencies, though the amplitude of the higher-frequency components coming within the short-wave band is much smaller than those in the long. Since noise is present at all frequencies, it follows that the wider the band-width of the receiver the more energy will it pick up from the atmospheric disturbance. When interference is of a continuous character, therefore, it is desirable that the band-

width should not be greater than that necessary to receive the signal. If, however, the interference takes the form of occasional pulses of large amplitude, then circuits of narrow band-width will distort and prolong the pulse considerably. In such cases the band-width may be made wide, so as to keep the pulse short, and a suppression circuit used which deadens the receiver for the very brief duration of the pulse.

External noise in the form of " man-made " interference has also to be considered in most situations and is worse the shorter the wavelength. All arrangements which produce sudden changes of current, especially if the change is accompanied by sparking, may produce the radiation of electromagnetic waves. Such waves will be heavily damped and may therefore be considered as a wide band of frequencies, the mean frequency (or frequencies) being determined by the effective inductance and capacity associated with the connecting leads to the apparatus producing the disturbance. Interference due to sparking brushes in a motor or similar causes can usually be reduced to inoffensive proportions at the source by the fitting of a condenser, or a condenser-choke combination across the motor terminals of sufficiently low H.F. reactance to make the production of an H.F. voltage practically impossible. The B.S.I. has issued specifications covering methods of measurement of interference (B.S. 727), components and suppression methods for various apparatus (B.S. 613), limits for interfering voltages and fields (B.S. 800) and the avoidance of interference at the receiver (B.S. 905). It is now a legal obligation to reduce radio interference from machines, etc. to reasonable proportions.

A powerful source of interference is that due to motor car and aircraft ignition equipment for the suppression of which there is no general legislation. This has the elements of a heavily-damped, spark transmitter and the electrical constants of the high-tension leads employed makes the interference most severe on wavelengths of about 6-7 metres. When wishing to use a receiver on any frequency in a car or aircraft, some form of suppression will be necessary. On aircraft engines this problem has been solved by the complete screening of all ignition leads right up to the engines and including the

sparking plugs and the bonding of the so-called screening harness according to a definite plan.

Interference may be caused by the insulators used on power transmission lines when such insulators are dirty or coated with salt, as then leakage currents flow across them. These are unsteady and probably travel by very small arcs between particles of deposit. The unsteady current is found to have components at radio frequencies and the overhead line acts as a good aerial to radiate the disturbance. Interference arising from the collectors of trolley-buses has been troublesome and filters incorporating suppressor coils carrying the supply current have been found necessary.

Interference can reach the receiver in a number of ways:

(*a*) By direct radiation from the offending machine to the aerial attached to the receiver, or to the receiver itself.

(*b*) The disturbance may produce R.F. currents in the mains supplying the machine and by stray mutual inductance or capacity couplings induce interfering E.M.F.'s in aerial or receiver.

(*c*) If the receiver is supplied from the same mains as the apparatus producing the disturbance, interference may be brought in along the lines.

(*d*) Even if disturbing apparatus and receiver are on different supply systems interference can be brought in by stray couplings between the two supplies.

The aerial can often be placed, with advantage, in a high, open position and connected to the receiver by a screened cable, the improved signal/noise ratio obtainable outweighing the attenuation introduced by the feeder. Clearly, the receiver itself must be well screened if the full benefit of this arrangement is to be derived.

On large ships, where there are a large number of electrical machines and it may be necessary to use radio transmitters whilst reception on other frequencies is going on, the whole receiving room is usually screened. Each power lead coming into the receiving room is fitted with a filter. To be effective these must be in a separate screened compartment. If inside the receiving room the disturbing currents flowing in the choke

of the filter may induce E.M.F.'s in the receiving circuits; if outside, and unscreened, the chokes may be the means of picking up interference and conveying it into the receiving room.

Internal Noise [3]

If all sources of external noise could be eliminated, there would still be noise generated in the receiver itself and thereby setting a limit to the weakness of incoming signal which can be handled if a satisfactory signal/noise ratio is to be obtained at the output. Assuming spurious noises due to faulty components, valves, batteries, etc., to be eliminated there still remain two fundamental sources of noise:

(*a*) Thermal-agitation.
(*b*) Shot-effect.

The electrons in the conductors of which a circuit is composed are all in random motion, the amplitude of motion being proportional to the temperature of the circuit. This motion is, of course, quite distinct from an ordered drift of electrons which take place if a p.d. is applied across the circuit. The motion of each electron constitutes a minute electric current which produces an E.M.F. across the ends of the resistance. As a result there are E.M.F.'s of all frequencies present across the resistance and the R.M.S. voltage measured across it will depend upon the band of frequencies accepted by the measuring apparatus. Since the R.M.S. value of a number of component voltages having different frequencies is given by

$$\sqrt{E_1^2 + E^2_2 + \ldots}$$

it follows that the R.M.S. voltage will be proportional to the square root of the band-width of the circuit.

It has been shown by Nyquist that the R.M.S. value of the E.M.F. produced in a resistance R is given by $E = 2\sqrt{kRT.\Delta f}$ where k is Boltzmann's constant (a physical constant relating absolute temperature and electron energy due to motion), T is the absolute temperature and Δf the pass-band of the measuring circuit (that is, the difference between the two frequencies at which the response is $\frac{1}{\sqrt{2}}$ of the maximum). This E.M.F. has to be taken as acting in series with R and

consequently, if a circuit is connected across R, the p.d. between its ends is less than E and can be calculated by the usual methods.

Suppose that we have a coil of R.F. resistance R and inductance L, tuned by a condenser of negligible losses. Then the E.M.F. generated in the resistance is $2\sqrt{kRT.\Delta f}$, and, by the usual relationship in a resonant circuit, this will provide a p.d. of $2Q\sqrt{kRT.\Delta f}$ across the condenser. But this may be re-written as $2\sqrt{\frac{\omega_o^2L^2}{R^2}RkT.\Delta f}$ which is $2\sqrt{R_e kT.\Delta f}$ where R_e is the effective parallel resistance or dynamic impedance of the circuit.

This relationship assumes that Δf, the band-width of the measuring apparatus, is smaller than the band-width of the resonant circuit which is being tested, but this is a practical condition, because the signal-frequency circuits in which the noise is most important are normally followed by much more selective circuits.

Boltzmann's constant is $1{\cdot}37 \times 10^{-23}$ joules per °C., and if we take 290° K as a usual room temperature, $2\sqrt{kT}$ is $12{\cdot}6 \times 10^{-11}$, so our expression for the E.M.F. generated in a resistor by thermal agitation becomes

$$E = 12{\cdot}6 \times 10^{-11}\sqrt{R.\Delta f}$$

and we have seen that this expression can also be used for the p.d. developed across a resonant circuit if R_e is substituted for R.

If we take as a typical case $R_e = 50{,}000\Omega$ and $\Delta f = 10$ kc/s, then the voltage developed across the tuned circuit by thermal agitation will be $2{\cdot}86\mu$V.

Such voltages are produced in all resistors in the receiver but, under normal conditions, only that produced across the input tuned circuit is of importance. If this is sufficiently below the level of the signal then, after the signal has been amplified, the noise produced in later circuits will be negligible compared with the signal.

The thermal-agitation voltage is used in some measuring apparatus as a standard voltage for adjusting the apparatus to give a standard amplification.

The shot effect is due to the fact that the electron current

in a valve is due to the movement of discrete electric charges and is not a perfectly continuous flow. As a result, the anode current fluctuates and a very small noise voltage is developed across the anode load. The current fluctuations contain all frequencies and the R.M.S. voltage across the load will evidently depend upon the frequency response curve of the load. The noise produced at the output of the receiver will, of course, depend upon the band-width of the succeeding circuits. In normal cases only the shot noise in the first valve will contribute appreciably to the final noise output.

The magnitude of the shot noise in a saturated diode can be calculated and diodes have been used as sources of known noise. A full quantitative treatment of shot effect in multi-electrode valves working under normal conditions is difficult, however, but we can arrive at some of the factors upon which its magnitude depends. We can see that it will be proportional to the anode current, whilst, on the other hand, the signal will be proportional to g_m. Hence it is desirable that the first valve in a sensitive receiver should draw a small anode current but have a large g_m, if the signal/noise ratio is to be good. It is also found that the partition of the emission current between the anode and the screens produces noise and, in consequence, the beam tetrode (in which the screen current is small) is the best multi-electrode valve from the noise point of view and the frequency-changers (see page 345) are the worst, because the transfer conductance is always much lower than g_m and screen currents are large.

It will be remembered that the use of a triode, with its grid earthed, as a power amplifier, was discussed in Chapter X. Special triodes, used in this way, have come into use for the first stage in very sensitive receivers, because a better signal/noise ratio can be obtained than with pentodes.

A useful way to specify the shot noise produced by a valve is to state the resistance which, placed across grid and cathode, would produce the same noise (due to thermal agitation) in the anode circuit, as does the shot noise. This enables a direct comparison to be made between input circuit noise and valve noise and allows for the g_m of the valve.

Typical values for this equivalent resistance are given in Table XXIII.

TABLE XXIII

Type of Valve.	Grid Resistance to Simulate Shot Effect.
Triode	200– 1,000Ω
Special beam tetrode. . .	800– 2,000Ω
Ordinary screen grid and pentode	5,000– 20,000Ω
Frequency-changer (Multi-grid) .	200,000–300,000Ω

Sensitivity of a Receiver

It is not easy to specify a set of measurements, not too tedious or difficult to carry out, which will enable the true merits of a receiver to be precisely stated. The Institute of Radio Engineers in the U.S.A. and the Radio Manufacturers' Association in this country have both issued a specification for a suitable set of measurements on receivers for amplitude modulation and particularly adapted to the broadcast receiver. Government departments such as the Post Office and Services usually evolve their own specifications and tests, suited for the particular type of receiver concerned. There are, however, two important measurements, sensitivity and selectivity, which are common to all receivers and which we will now discuss.

By the sensitivity of a receiver is meant its maximum sensitivity, and may be defined as the minimum input, specified as below, which will give a certain defined output power in a non-reactive load resistance of the same magnitude as the impedance of the loud speaker, or other output, under a prescribed condition of signal/noise. For receivers intended to work with loud speakers, 50 mW is usually taken as the output power but 1 mW may be used for receivers intended for phone or line work. The signal/noise ratio prescribed is determined by the type of receiver involved. For receivers working at a high field level, as broadcast receivers, 15 db is usually specified, but with receivers operating at very low field levels, as communication and commercial receivers, 10 db is a more usual figure.

One common method of carrying out a sensitivity test is as follows: A signal generator is connected, through an artificial aerial of specified constants, to the input of the receiver. An output meter is connected across the output, in place of loud

speaker, or other output. The output meter comprises a variable resistive load (to match various output circuits) and a voltmeter which measures the voltage across the load.

The signal generator is then set to give the required carrier frequency, modulated 30% at 400 c/s. The gain of the receiver is adjusted to give full output. The modulation is then switched off and the output of the receiver, which is now due only to noise, is measured on the output meter. The difference in output from the two conditions evidently gives the signal/noise ratio. If this is less than the prescribed value (say 15 db) then the test must be repeated with a larger input from the signal generator. If the ratio is higher than required a smaller input can be tried.

Selectivity

The selectivity of a receiver concerns its ability to discriminate between the desired signal and those at other frequencies without distorting signals within the pass-band. If we refer to Fig. 337, the selectivity may be specified in the following manner: state the pass-band (f_p) and the variation of level α_p db, which may be permitted within it. Specify the loss in db (α_c) at a frequency (f_c) slightly outside the pass-band. The greater the difference between α_c and α_p and the less the difference between f_c and f_p, the better will be the circuit characteristic. For example, in a commercial telephone receiver the values will be about as follows: $f_p = 5$ kc/s, $\alpha_p = 6$ db, $f_c = 7{\cdot}5$ kc/s, $\alpha_c = 80$ db.

In the simpler classes of receiver we would have to be content with a less elaborate type of circuit, giving inferior selectivity, as shown dotted in Fig. 337.

These curves fail to take account for the effects due to the fact that, when using the receiver, wanted and unwanted signals are applied at the same time. Suppose that there is a powerful, unwanted signal at the input, then the later tuned circuits in the receiver may be capable of reducing it to a harmless value. If, however, the early valve circuits are not quite linear in operation then cross-modulation may occur—that is, the modulation of the unwanted signal will become superimposed on the wanted carrier and will therefore reach the detector in spite of the filtering action of the later tuned

circuits. The interfering signal will therefore be more troublesome than the frequency/response curve would indicate. This effect may be particularly important in receivers which have to work near transmitters (on board ship, for example). In such cases it may be necessary to have a number of tuned circuits preceding the first valve.

If the circuit is non-linear then an unwanted signal which may be sufficiently far removed in frequency that its carrier does not come through can still produce interference if a

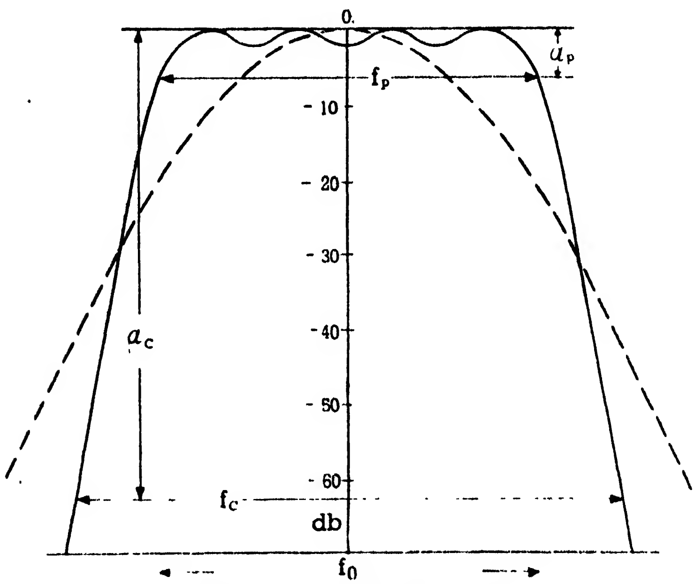


FIG. 337. Illustrating Selectivity.

portion of its side-band only is received. In this case the side-band will beat with the wanted carrier and produce interfering frequencies. For example, an unwanted carrier, 14 kc/s off-tune, applied to the receiver to which the response curve of Fig. 337 applies, would not in itself be troublesome, but if it is modulated by sinusoidal frequencies of 5 and 10 kc/s, the upper side-bands which these frequencies produce will reach the detector. These will beat with the wanted carrier to produce frequencies of 9 and 4 kc/s.

Another effect which modifies the selectivity of the receiver

is the demodulation* effect. If an unwanted R.F. voltage reaches the detector which is receiving a desired modulated signal, then the amplitude of modulation is reduced. This action can be understood by considering the diode detector. The rectification of the unwanted carrier produces the same effect on the detector as a fixed negative bias.

If the interfering signal is also modulated, the demodulation effect will be mutual, that is, the desired carrier will reduce the interfering modulation. Since the demodulating effect is proportional to the magnitude of the demodulating carrier, if the interfering carrier is weaker than the desired one, then the interfering output at modulation frequency is reduced by the desired carrier to a greater extent than the interfering carrier reduces the wanted output and the performance of the receiver is improved.

An ideal receiver should yield an output at the modulation frequency which has the same waveform as the envelope of the modulated radio-frequency put in from the aerial. This requires that: (*a*) All frequencies involved in the signal should be amplified to an equal extent; (*b*) the phase shift should either be zero or proportional to the frequency; (*c*) the output should be strictly proportional to the depth of modulation in the input, that is, the receiver should be operating in a linear manner.

In practice, when receivers are being employed for telephony, (*b*) is unimportant but must be considered in television or radar receivers. To satisfy (*a*) the overall response curve must be sufficiently broad and flat, whilst (*c*) requires careful design of detector and low-frequency amplifiers. Non-linearity of radio-frequency stages produces distortion of the output when cross-modulation occurs, as has already been discussed.

Control of Receivers

Apart from a commercial receiver, which may remain tuned to the same station for hours at a time, and where the high performance justifies some elaboration of manual controls, the average receiver should have as few manual controls as possible, particularly so in the case of a broadcast receiver.

Where automatic controls replace manual they should give

* "Demodulation" here means the reduction of amplitude modulation. It is frequently used as an alternative to "detection."

results at least equal to those which can be achieved by a skilled operator. Automatic gain control (also termed automatic volume control), that is, the control of the overall gain of a telephone receiver so as to keep the output constant when the input varies, is greatly superior to any manual control of volume. Other forms of automatic control which are coming into wider use are : automatic frequency control which aims at keeping the receiver in tune ; and automatic selectivity control which is designed to keep the pass-band to the minimum width necessary for the signal being received.

Valves, Circuits and Components at Very High Frequencies

The general problems concerned with using valves at very high frequencies have been discussed when dealing with power amplifiers (Chapter X). From the receiver point of view, the decrease in the input resistance brought about by cathode-lead inductance and transit time is the most serious feature.

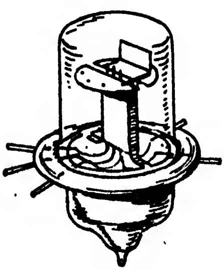


FIG. 338. Acorn Valve.

The "acorn" valve was one of the first valves suitable for very high frequencies, being shown full size in Fig. 338. This valve was difficult to manufacture and correspondingly expensive. It introduced the ring-type seal, made by pressing the edges of two cup-shaped glass envelopes together, the lead-in wires being disposed fan-wise round the edge of these cups.

Later, other types, such as the EF50, employing the ring seal in a somewhat different way, were developed. These were not so small but were suitable for very high frequencies, and were manufactured in enormous quantities during the war.

As the frequency is increased into the ultra-short waveband, great care is needed in the layout of the circuits, since even short lengths of lead may introduce large, series inductive-reactances and small, shunt capacitive-reactances. Screening will need to be thorough since the coupling between short lengths of lead may be sufficient to give rise to a significant amount of positive or negative feed-back. The earth connec-

tions belonging to one stage are frequently taken to a common point (provided that this does not involve long leads) instead of using the chassis as a connection, as is usual at lower frequencies.

Components must be carefully chosen. It is important to realise that all components have the three properties of resistance, inductance and capacitance. Thus we may call a component a resistance because, at low frequencies at least, the resistance quite swamps reactive effects. As we raise the frequency, however, reactance may greatly modify the performance of our " resistance." A little consideration will show that series inductance will have most effect in the case of low resistances, whilst shunt capacitance influences high resistances more.

Any wire-wound resistance (even non-inductively wound) will show pronounced reactive effects at 50 Mc/s, but the lower-valued composition resistances are still practically non-reactive at 100 Mc/s.

A condenser has an effective shunt resistance because of dielectric loss and hence has a power factor. The power factors of some materials are listed on page 411. The leads and plates of a condenser also represent a series inductance so that the component can be represented by an equivalent circuit such as Fig. 339.

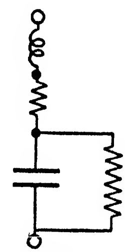


FIG. 339. Equivalent Circuit of a Condenser.

At low frequencies the inductive reactance is quite negligible but as the frequency is raised it increases and neutralises some of the capacity reactance so that the apparent capacitance of the condenser rises. Eventually the circuit resonates and at higher frequencies than this the condenser behaves as an inductance.

A series of measurements by Dr. Hartshorn [4] at the N.P.L. showed that at 50 Mc/s, a 0·0005 μF mica condenser had a lower impedance than a 0·1 μF so-called " non-inductive " tubular condenser although in both cases the condenser was inductive. Smaller condensers (100 $\mu\mu$F) showed capacitive reactance, the apparent capacitance increasing with frequency.

It might be expected that decoupling would be comparatively easy at high frequencies, since even a small capacitance will

have a low reactance, but the above results show that care is necessary in the selection of capacitors for the purpose. The average paper capacitor is either of the rolled tinfoil type or a rolled paper strip sprayed on one side with a conducting material. Some types are quite useless for the purpose of shunting high frequencies.

Condensers made with a dielectric using a ceramic material of large dielectric constant can be made in such small dimensions that inductive effects are negligible, and such condensers are particularly suited to high-frequency work.

Coils have an effective self-capacitance between their terminals as well as an effective series resistance due to the various losses. In consequence, as the frequency is raised, a coil comes into parallel resonance and, above this frequency, behaves as a capacitor. If the Q of a coil is fairly low, therefore, it will exhibit a flat resonance curve and can be used as a radio-frequency choke, giving a high impedance over a range of frequencies.

A Tuned Radio-Frequency or "Straight" Receiver

A typical receiver of this type is shown in Fig. 340, the circuit being that of a receiver which has been used in the mercantile marine. On the short-wave bands the R.F. amplifier will not add much to the overall sensitivity or selectivity of the receiver but serves to isolate the detector from the aerial. This partially prevents radiation when the detector is made to oscillate and also prevents changes in the aerial circuit affecting the frequency of oscillation. The Amplifier Gain Control varies the bias on the variable-μ R.F. valve, this bias being derived from the valve feeds.

In a short-wave receiver of this type, reaction in the detector circuit is essential if adequate sensitivity is to be obtained. The receiver will be most sensitive for telephony reception when it is just free from oscillation, whilst for C.W. telegraphy it must be oscillating in order to produce the beat note, but the oscillations should not be too large or the detecting action is impaired. A smooth control of reaction is very important. In this receiver the control of the screen-grid voltage, marked Detector Gain, will also control reaction, in addition to the reaction condenser.

A receiver of this kind will often employ plug-in coils and, if used for short waves only, will have tuning condensers of about 100 $\mu\mu$F maximum value. Although this small size means that more coil ranges are required, it enables the impedance to be kept higher and less variable over each range. The smaller condenser also results in frequencies being more spread out on the tuning dial.

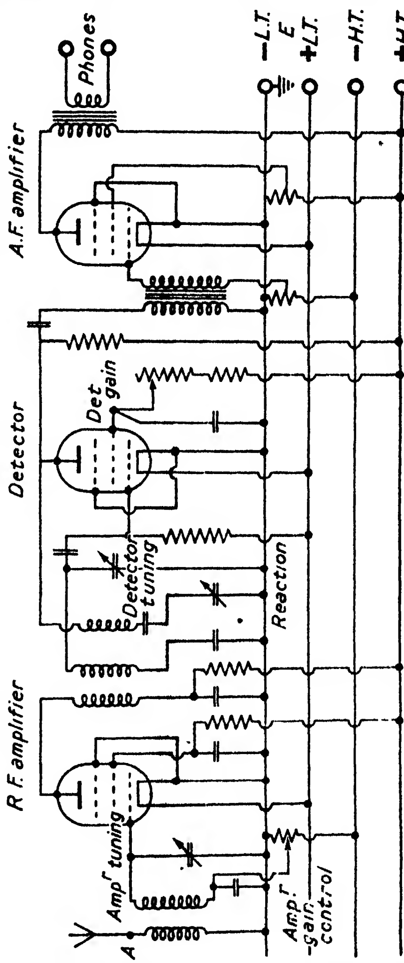


FIG. 340. Typical "Straight" Receiver.

The detector adjustments will be critical if good results are to be obtained. The amplifier tuning will be flat, since the Q of this circuit, damped by the aerial and valve, will be low, but the detector tuning will, of course, be very sharp when the reaction has been brought up. Since the response curve will then be a very sharp peak, it is evident that the quality of reproduction of telephony cannot be very good.

The method of receiving telegraphy by making the detector self-oscillate is unsatisfactory on long waves because the detector circuit is then too far out of tune but on short waves the percentage mistuning to give an audible beat note is, of course, negligible.

The disadvantages of the "straight" receiver, particularly for short and ultra-short wave reception, may be summarised thus:—

(1) Each tuned circuit needs its own variable condenser and

difficulties of ganging limit the number of stages, where simplicity of controls is wanted.

(2) The amplification of any stage is dependent on the resonant parallel-resistance (L/CR) of the tuned circuit in the anode of each wave. Owing to the large variation of C over the tuning range, R_e varies considerably. Thus the sensitivity is low at the low-frequency end of each range, increasing to the high frequency.

(3) Attenuation at a given off-tune frequency is proportional to Q/f. If there is no reaction, Q changes but little so that attenuation is inversely proportional to f and selectivity falls as the carrier frequency rises.

(4) Especially on ultra-short waves it is difficult to get adequate amplification because of the valve defects previously discussed and low Q circuits. If reaction is resorted to, in order to boost amplification, the quality of telephony deteriorates and the receiver requires critical adjustment.

The advantages are :—

(1) Simplicity, especially in the general-purpose type of receiver described above, for use by skilled operators.

(2) Spurious signals cannot be produced in the receiver itself.

After considering the disadvantages, it seems surprising that some successful television sets employ "straight" receivers, and they were also used on some of the metre-wave radar sets. In both cases, however, the receiver works on a fixed (or preset) tuning and this simplifies the design and testing enormously, enables a compact receiver to be built and the various stages to be more thoroughly screened. The band-width required is very great and hence the low Q of the circuits is no handicap—in fact, the Q has frequently to be artificially lowered by shunt resistors. Amplifiers working on a frequency of 45 Mc/s with a band-width of 4 Mc/s have been designed (employing valves such as the EF50) which give a voltage gain of about 10 times per stage. By very careful attention to design, five or six stages have been used without instability.

The Super-Regenerative Receiver

We have already seen, when discussing the detector with reaction usually used in a straight receiver, that very considerable amplification can be obtained, since reaction can make the

series resistance of the circuit very small (or the conductance very large). Unless we wish to use heterodyne reception, however, it is not possible to make the conductance infinite because oscillations will then commence.

Armstrong showed that if the conductance of a circuit can be varied periodically (at a frequency considerably lower than that of the incoming frequency) between positive and negative valves, then a very large amplification of a modulated signal can be obtained. He termed the process " super-regeneration."

The usual method of obtaining the variation in conductance is to add an alternating voltage to the anode supply voltage of the valve producing the reaction. This alternating voltage

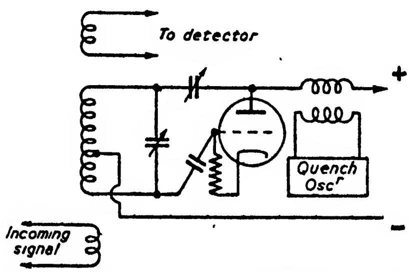


FIG. 341. Principle of Super-regenerative Receiver.

is normally produced by an auxiliary oscillator, called the " quench oscillator."

Thus the basis of a super-regenerative receiver might be as in Fig. 341. By means of the condenser between anode and tuned circuit, the circuit would be adjusted so that it has negative conductance for at least a part of the half-cycle during which the quench oscillator is adding to the anode voltage, but has a positive conductance during the remainder, a possible variation of conductance being shown in Fig. 342.

An incoming signal will start oscillations at the resonant frequency, f_0, of the circuit. If the circuit is so adjusted that the quench oscillator stops the growth of oscillations before their amplitude becomes limited by the valve characteristics, it is said to be working in the linear mode. In this case the oscillations form a pulse, as shown in Fig. 342—growing all the time that the conductance is negative, so that they reach their maximum amplitude at the end of the negative conductance

period and then decay again. The pulses evidently repeat at the quench frequency, f_q. The amplitude reached by the pulses is proportional to the value of the signal at the moment oscillations started to build up.

An alternative mode of operation is when the oscillations are allowed to build up until limited by curvature of valve characteristic, before the quench oscillator causes them to decay. In this case all signals above a certain minimum value will produce the same amplitude of pulse, but it is found that the area under the pulse envelope is proportional to the logarithm

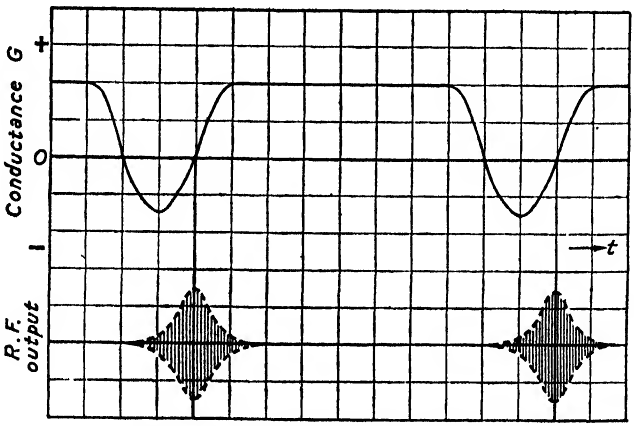


FIG. 342. Illustrating Quenching Action.

of the signal amplitude, and this type of operation is therefore referred to as the logarithmic mode.

We have seen that pulses of radio frequency such as shown in Fig. 342 may be resolved into components at f_o, $f_o \pm f_q$, $f_o \pm 2f_q$ having relative amplitudes which depend on the shape of the pulse envelope.

Now suppose that the incoming signal is modulated. Provided that f_q is considerably higher than the modulation frequency, the amplitude of the pulses will follow the modulation. As the pulses are now varying in amplitude, each component frequency will be accompanied by a pair of side-bands. It is necessary for f_q to be several times greater than f_m if these side-bands are not to interfere with each other and produce distortion in the rectified output.

It can be shown that the gain produced, when working in the linear mode, is proportional to the negative conductance produced and to $1/C$, where C is the capacitance in the circuit. This gain may be much larger than that which would be produced if the conductance was reduced to a low value without going negative, as in an ordinary circuit with reaction. The fact that gain is proportional to $1/C$ makes the super-regenerative receiver specially suitable for ultra-short waves.

Clearly noise voltages due to thermal agitation and shot effect will also initiate oscillations and produce pulses at f_a. These would be inaudible, if all of the same amplitude, but they

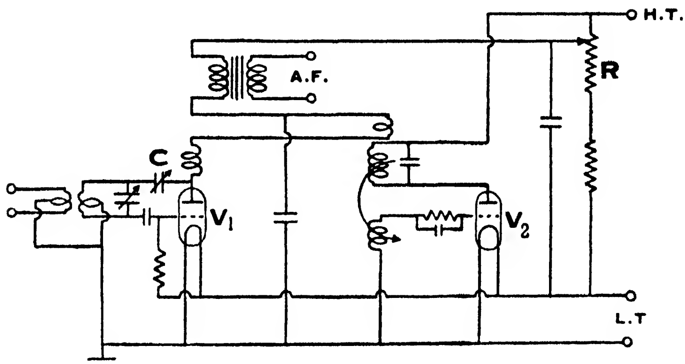


Fig. 343.

will be fluctuating as the initiating voltage fluctuates. Hence a noise is produced in the output, which may be considerable.

External noise coming in with the signal evidently alters the value of the total signal at the instant when oscillations commence to build up, and hence causes the amplitude of the pulses to vary, that is, modulation additional to that in the signal is introduced, so yielding noise at the output.

It is evidently possible to carry out detection in the oscillating stage, in the same way as in the simple, straight receiver. A typical super-regenerative receiver for ultra-short waves is shown in Fig. 343. The aerial is connected through a screened feeder to the closed LC circuit by means of a variable mutual coupling, the adjustment of the latter providing a sensitive control of signal strength. V_1 is the super-regenerative

detector and V_2 the quenching valve. The grid-leak of the detector is returned to the positive side of the filament and the Hartley-type oscillator is controlled by the condenser C and reaction by R. This type of circuit is simple and free from spurious responses, which may cause squeggers. For the best results, and to keep the noise-level down, the coil should have a high Q, so that minimum reaction is necessary. The quenching voltage is injected into the anode circuit rather than the grid, as this has been found to give more stable operation. The quenching circuit oscillates at a frequency of about 20 kc/s, a value low enough not to reduce sensitivity but high enough to allow of easy filtering in the audio-frequency stage.

It is of interest to compare the reaction detector and the super-regenerative circuit, since both form sensitive receivers using few valves and capable of being made very compact. The gain of the super-regenerative receiver can be considerably greater, though its signal/noise ratio will be poorer. In neither case is high-quality reproduction of a modulated signal likely to be achieved.

The super-regenerative receiver is not selective because negative conductance has the same effect as positive, as far as flattening out the resonance curve is concerned (since it is a component of the impedance which does not depend upon frequency). For very short-wave applications this lack of selectivity is an advantage, since the available frequency band is so great and many transmitters are not very stable. The reaction detector, on the other hand, is very selective when adjusted for large gain.

The controls of the super-regenerative receiver are not critical, once a suitable circuit has been set up. Both tuning and reaction controls are very critical in the reaction detector, if best results are to be obtained.

The super-regenerative receiver has generally been considered to be somewhat unreliable but during the war thoroughly reliable types have been produced in great numbers, to operate on about 200 Mc/s. These were carried on aircraft for the purpose of receiving radar pulses and automatically causing them to send out pulses from a transmitter. The signals from this transmitter, being received at the radar station, identified the aircraft as a friendly one.

A compact, yet very reliable receiver, was evidently required and the super-regenerative receiver, working in the linear mode, with a quench frequency of about 300 kc/s, was found very suitable. A gain-stabilisation circuit was incorporated.

The Super-Heterodyne Receiver [1]

This system has the advantage that nearly all the amplification may be provided at a fixed frequency, thereby reducing adjustments and simplifying the design of the amplifying circuits. The band-width of the receiver may also be kept

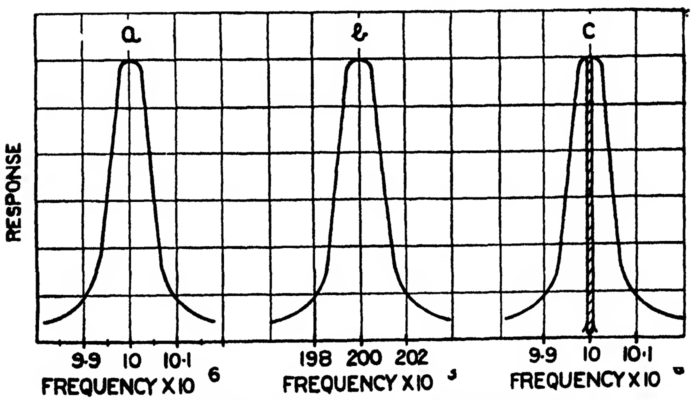


FIG. 344. The Selectivity of Super-heterodyne Receivers.

the same over the whole wave-range for which the receiver is designed, this being difficult in the "straight" receiver.

In the super-heterodyne, the desired carrier frequency, together with a local-oscillator frequency, is applied to a frequency-changer valve which produces in its anode circuit components of the sum and difference frequencies of the two oscillations. Either of these frequencies, which carry the modulation, may be selected by suitable filters, although the difference-frequency is always chosen, and amplification and selection after the frequency-changer will thus be carried out at a fixed frequency known as the intermediate frequency (I.F.).

A consideration of Fig. 344 will show that the same selectivity can be imparted to a receiver by one or two stages tuned to the I.F. as by many tuned to the radio frequency (R.F.).

Curves (*a*) and (*b*) are for two circuits having equal damping but tuned to a R.F. of 10 Mc/s and an I.F. of 200 kc/s respectively. When the curves are combined it is seen the overall selectivity is very high.

Sufficient tuning at the R.F. must be provided, of course, to ensure that the "image frequency" or "second channel" interference is cut out; that is, an unwanted frequency having a frequency differing from that of the wanted carrier by twice the I.F. must be rejected before passing to the frequency-changer, which would be unable to differentiate between it and a wanted signal.

The choice of I.F. to use in a receiver is governed by the following considerations. A low I.F. makes it easier to produce

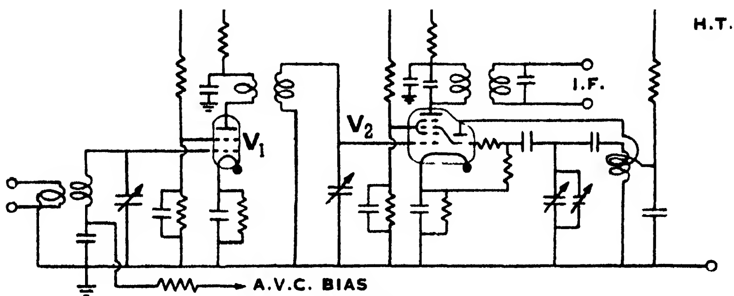


FIG. 345.

a very selective receiver but more R.F. tuning is required to eliminate the image frequency, whilst a very high I.F. will make the provision of the necessary amplification more difficult. The use of a low I.F. makes it more difficult to prevent interaction between the beating oscillator and the signal tuned circuit, and for a general purpose circuit an I.F. of about 450 kc/s is usual.

In a frequency-changing system interference can be caused by the interaction of the harmonics of the frequencies beating together, owing to the fact that the frequency-changing valve must have asymmetrical characteristics and therefore produces harmonic distortion. Such interaction (which may result in audible whistles) is reduced by the first circuit tuning, by a careful choice of the I.F. and by not allowing the oscillator amplitude to be too large.

The super-heterodyne lends itself to automatic control of volume, tuning and selectivity. It is, therefore, in many ways superior to the other types of receiver, its only undesirable feature being that just mentioned, namely, its ability to produce unwanted audio frequencies through the frequency-changing system.

Some complex super-heterodyne receivers will be described in the next chapter, but in Fig. 345 the radio-frequency and frequency-changing circuits of a typical, general-purpose super-heterodyne are shown.

It is very desirable to have at least one stage of R.F. amplification, as well as tuning, because (as already mentioned) the valve noise of frequency-changers is high. In the circuit shown, an aligned-grid tetrode is employed because such valves have a specially-low equivalent noise resistance.

One of the difficulties in designing a super-heterodyne for short waves is that at high radio-frequencies the ratio between oscillator and incoming frequencies decreases, and interaction between the two circuits becomes greater. This may be due to inter-electrode capacity or electron stream coupling. This has two effects ; it causes the R.F. circuit to reflect a resistance and reactance component across the oscillator circuit, the first affecting oscillator amplitude, the second oscillator tuning ; it also causes an oscillator voltage component to appear across the R.F. circuit, and as this is usually out of phase with the oscillator-electrode voltage it reduces the effective oscillator voltage applied to the frequency changer. In the heptode frequency-changing valve, where the oscillator electrode is nearest the cathode, electron coupling has the same kind of effect but it is opposite in direction to that exercised by inter-electrode capacitance.

At one particular frequency and set of operating conditions electron coupling may be completely cancelled by capacitance coupling and a condenser of 1 $\mu\mu F$ is sometimes connected externally between signal-frequency and oscillator electrodes to reduce this coupling effect over a range. In the hexode type of valve, this type of electron coupling is almost absent because the control grid is nearest the cathode, but there is another effect which must be allowed for. Here the oscillator voltage repels some of the electrons with sufficient velocity to

be collected by the control grid when its voltage has the least negative value. The frequency for which favourable conditions for the collection of electrons arises increases with increase of signal-frequency, so that a bias of about $-2 \cdot 5$ volts may be necessary at 20 Mc/s to prevent grid current whereas only about $-0 \cdot 5$ volt is necessary at 1 Mc/s.

The selectivity of any radio-frequency tuned circuit following the first valve should be high, not only to remove the image signal but to prevent noise side-bands and undesired signals beyond audio range of the desired carrier from reaching the frequency-changer, where they can combine with the oscillator or its harmonics to produce interference frequencies.

For the oscillator, either a Hartley or tuned-grid circuit is used. Such circuits are easier to maintain in oscillation over a range of frequencies than a tuned-anode circuit and they give a more constant output over the range.

The oscillator section of a triode-hexode has a low g_m and in the tuned-grid circuit tight coupling is essential in order to ensure that oscillation is maintained over the whole range. It is usual to interleave the turns of feed-back and tuning coils.

The tight coupling may cause a tendency to squegger at the high-frequency end of the range but this can usually be overcome by using a low value of grid leak (not greater than 50,000Ω) and grid condenser (not greater than 50 $\mu\mu F$) and a resistance of 50Ω is often included in series with the grid to reduce oscillation amplitude at high frequencies.

In the more elaborate receivers a separate oscillator valve is often fitted because its frequency stability will usually be better and also because the low g_m of the triode-hexode makes it liable to stop oscillating if the heater supply volts fall to a value which is not low enough to put the rest of the receiver out of action.

Mixer for Centimetric Waves

When frequencies such as 10,000 Mc/s are being employed, no amplification at the original frequency is possible. External noise is low and therefore the smallest signal which can be made use of is largely determined by the internal noise level. It is possible to use klystron amplifiers at these frequencies but the

signal/noise ratio is found to be poorer than if the incoming signal is taken straight to the frequency-changer.

It has already been noted that frequency-changer valves introduce more noise than other types and that this is one of the main reasons why amplification at the signal frequency is used wherever possible. If we have to dispense with this at 10,000 Mc/s, then one of the most important factors which will influence our choice of frequency-changer will be the amount of noise which it introduces.

The crystal detector has been found better than any valve, from the noise point of view. Since valves will have a very low input impedance compared with that at lower frequencies, the sensitivity of the crystal is not very different to that of the valve.

The crystal detector, of the "cat's whisker" type, has therefore come into use again but in a greatly improved form. Instead of requiring continual adjustment in use and giving quite unpredictable results, the modern crystal detector is permanently adjusted after manufacture and quite stringent acceptance tests are applied.

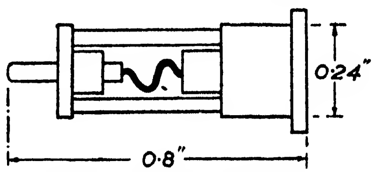


FIG. 346. Detector for Centimetric Waves.

A section through such a detector is shown in Fig. 346. The crystal is silicon and the contact-wire tungsten. The silicon and the wire are carried by brass ends which are firmly fixed in a rigid ceramic tube. A sensitive point of contact is found and then the capsule is filled with wax, which prevents movement of the contact and also stops moisture from getting in.

Pure silicon has not been found suitable and much experimental work has been done to find the nature and exact quantities of impurities which should be added. Aluminium, beryllium and boron have all been used.

When using the detectors for radar receivers, they have to be capable of withstanding, for very brief periods, high voltages from the transmitter and the impurities have a considerable influence on their ruggedness in this respect.

When a crystal is used as a mixer, it is found that there is an optimum value for the local oscillator voltage applied to it.

An increase of oscillator voltage at first sets the crystal to a better rectifying point and therefore increases the signal output. At the same time, however, the noise generated in the resistance of the crystal rises and eventually leads to a poorer signal/noise ratio.

The capacitance of the contact is of the order of 0·5 $\mu\mu F$ and this acts as a shunt to the rectifying action. The constants of a typical crystal are such that the performance of the crystal is much the same at all frequencies up to about 3,000 Mc/s but between this and 10,000 Mc/s there is a marked reduction in the rectified output for a given input. The klystron is usually employed as the local oscillator.

It has long been known that the B–K electron oscillator can be used as an oscillating detector for centimetric waves but it has been superseded, in modern equipment, by the crystal mixer and klystron oscillator.

Reception of Frequency-Modulated Waves

F.–M. is used on ultra-short waves for communication and for high-quality broadcasting. In the former case the deviation

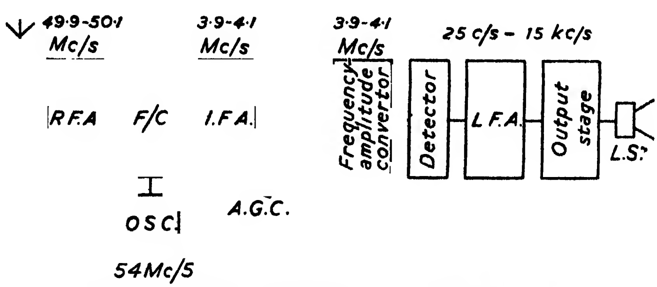


Fig. 347. Schematic Diagram of F.–M. Receiver.

frequency may be about 20 kc/s but in the latter case may be as high as 100 kc/s.

A super-heterodyne receiver will normally be employed and Fig. 347 shows in block form the various circuits. The earlier stages will be much the same as for an A.–M. receiver working on the same frequency. If wide deviation is being employed, a high I.F. will be necessary, about 4 or 5 Mc/s being suitable. Considerable I.F. amplification is required, because the detecting arrangements need a large signal if they are to function

well, and much amplification at the original frequency is not possible on ultra-short waves.

It will be necessary to take precautions to make the frequency stability of the oscillator in the frequency-changer very good. A slow drift of frequency will seriously upset the working of the detecting arrangements whilst rapid fluctuations, such as might be produced by a ripple on the anode supply to the oscillator, would cause frequency modulation and therefore noise output.

The Limiter

We have already seen that one great advantage of frequency modulation is the improved signal/noise ratio which is possible. This can only be obtained, however, if we prevent any ampli-

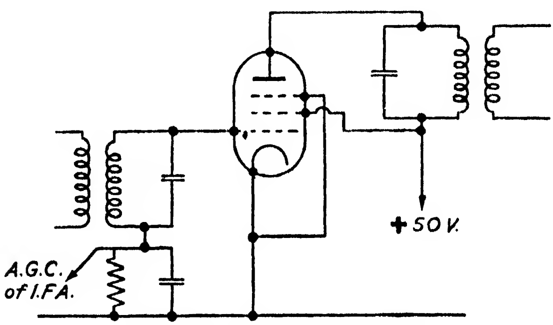


FIG. 348. Circuit of Limiter.

tude modulation, impressed upon the input signal by interference, from yielding an output, that is, the detecting arrangements must be very insensitive to A.-M. This is accomplished by inserting a limiter, the most common form being shown in Fig. 348.

There being no grid bias, grid current flows and builds up a steady voltage across the grid leak, exactly as in the case of a leaky-grid detector. This grid bias is clearly proportional to the signal amplitude. As the valve is worked with small anode and screen voltages, the positive peaks of comparatively small signal voltages produce anode currents up to the saturation value. The negative peaks produce cut-off. Hence larger signals do not give larger outputs. A typical curve connecting input and output voltage is shown in Fig. 349.

The time-constant of the grid-leak circuit must be small

enough to allow the bias to follow the amplitude modulation. The limiting action clearly distorts the R.F. waveform but this is of no consequence as long as the tuned circuits which follow are capable of cutting out the second harmonics produced.

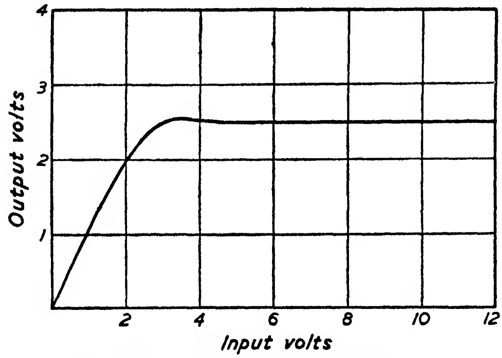


FIG. 349. Characteristic of Limiter.

The voltage produced across the grid leak of the limiter can be conveniently used for automatic gain control of the I.F. amplifier.

A number of different frequency-amplitude converters have been used, but we shall discuss only the one which is probably the best and the most used.

The Phase Discriminator

A circuit diagram is shown in Fig. 350, whilst its behaviour can be understood from the vector diagram of Fig. 351. It

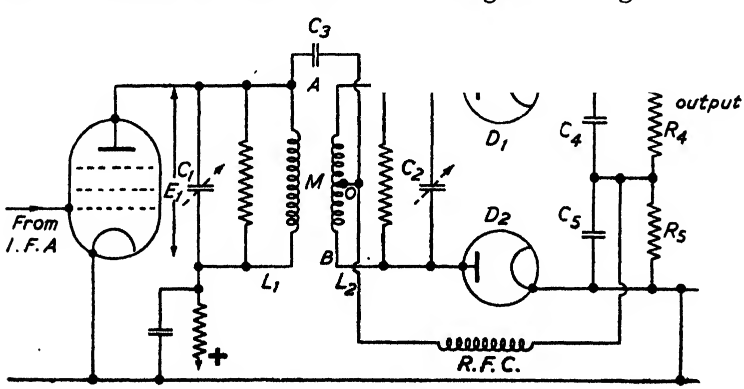


FIG. 350. Circuit of Phase Discriminator.

will be seen that E_1, the R.F. voltage developed in the preceding circuit, is applied through C_3 and O to both D_1 and D_2. Added to these voltages are those present across OA and OB due to the mutual inductance coupling between the two resonant circuits. The connection between the junction of R_4 and R_5 carries only rectified current from the diodes, owing to the presence of the choke.

The full lines of the vector diagram (Fig. 351) show the resultant voltages E_{D1} and E_{D2} applied to each diode when the frequency is equal to the resonant frequency of the circuit. In this diagram E_{OM} is the voltage between O and A, whilst E_{OP} is that between O and B.

The quadrature relationship between E_1 and E_{OM} or E_{OP} arises in the following way: E_1 produces a current in L_1 given by $\dfrac{E_1}{j\omega L_1}$ and this, in turn, induces an E.M.F. $-\dfrac{j\omega M E_1}{j\omega L_1}$ or $-\dfrac{ME_1}{L_1}$ in L_2. Since the secondary circuit is in resonance, the current in it will be $-\dfrac{ME_1}{L_1 R_2}$ (where R_2 is the resistance of the circuit) and $E_{OM} = -\dfrac{j\omega L_2 M E_1}{2L_1 R_2}$, which is in quadrature of E_1. It will be seen that E_{D1} and E_{D2} are of the same magnitude and will, therefore, produce equal rectified voltages across C_4 and C_5 but of opposite polarity and the L.F. output is zero.

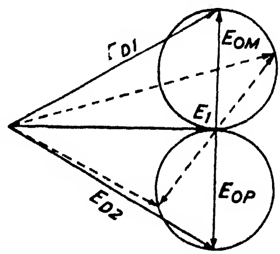


FIG. 351. Vector Diagram for Phase Discriminator.

If the frequency of the voltage injected in series into any resonant circuit varies, the vector of the current in the circuit moves so that the locus of its end is a circle having the current at resonance as its diameter. It follows that the locus of the voltage across the circuit will also be very nearly circular. When the frequency is deviated, therefore, conditions will become as shown by the dotted-line vectors. It is here assumed that E_1 does not shift in magnitude or phase due to the change in frequency. It will be seen that different voltages are applied to D_1 and D_2 and hence an output appears.

The magnitude of this output (for a given frequency devia-

tion) would depend upon the magnitude of the input voltage and the limiter is necessary in order that the arrangement may not respond to amplitude modulation, produced by noise.

The shunt resistances in the resonant circuits are to reduce the Q to the low value necessary if the required deviation is to

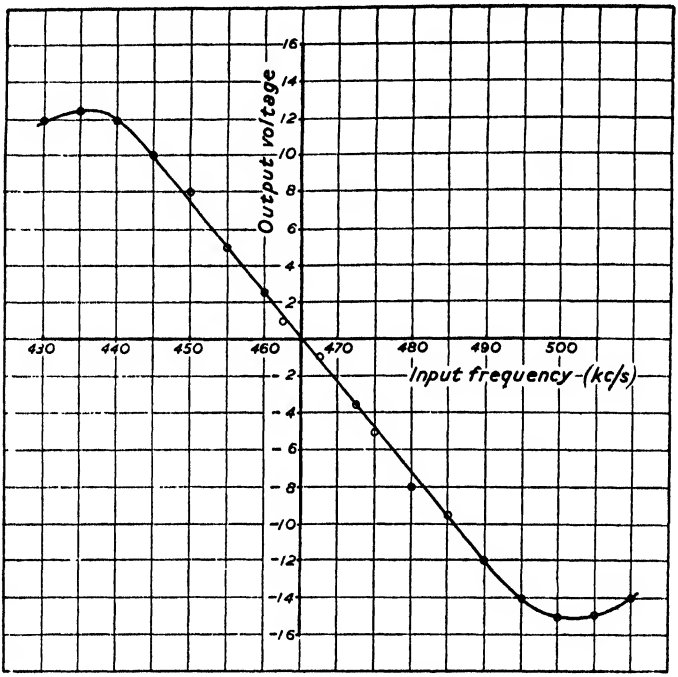


FIG. 352. Performance of Discriminator.
I.F. = 465 kc/s ; $\Delta f = \pm$ 15 kc/s ;
Q of each circuit = 12.
Coeff. of coupling = 0·12.
L_1 = 653 μH ; L_2 = 1,160 μH.
C_3 = 68 $\mu\mu F$; $C_4 = C_5$ = 150 $\mu\mu F$.
$R_4 = R_5$ = 0·1 MΩ.

be accommodated. Fig. 352 shows the performance of a certain discriminator having the given circuit constants.

Diversity Reception

It has been found, experimentally, when recording high speed telegraph signals, or telephony, that the sudden " deep

fades " that are so troublesome are very local; that is, when the signal from a station is at a very low level in one spot at a certain instant, it may be quite high in another spot only a few hundred feet away.

In consequence of this fact, considerable work has been done on combining the signal received from several aerials so as to produce a more uniform output level and eliminate as many deep fades as possible.

The problem of " mixing " the E.M.F.s given by the various aerials is quite different from that encountered when spaced aerials are used to give directional properties, for in this case the relative phases of each aerial must be retained, as these determine the directive properties. In diversity reception we wish to combine in a scalar fashion the outputs from each aerial (or array), irrespective of the phase differences which must exist due to spacing. There are at least three possible schemes.

(1) Use a separate receiver for each array and combine their low frequency outputs. The disadvantage of this method is that when a deep fade occurs on the array the receiver connected to it gives no signal output, but still gives some noise output, and therefore the resultant signal/noise ratio suffers. Another disadvantage is that if any of the arrays are receiving a very distorted signal, due to the selective fading of the carrier frequency at any instant, then this array will contribute distortion to the final result.

(2) Use a separate receiver for each array as before, but automatically select the signal mainly from the receiver giving the best output. This evidently avoids the faults of (1).

(3) After some separate amplification, use a rotating switch (or an electronic equivalent), which will apply the arrays in rapid succession to the remainder of the receiver. In order to be effective, this switching must connect each array at least once during each Morse dot, and hence is really modulating the signal. This necessitates a wider frequency band in the receiver, and this in turn increases the noise level.

It is generally agreed that the use of some diversity method of reception is desirable, but it does not, as is often supposed, provide a simple and cheap system.

Not only must two or more receivers per circuit be provided,

but if a good service is required, the diversity aerial cannot be of the simple omni-directive type, but each must comprise in itself a small array.

The Radio Corporation of America has used the diversity principle widely, three receivers being employed and combining according to either (1) or (2) above.

Diversity reception has also been brought into considerable use by Cable and Wireless and at the Somerton station a long series of tests was undertaken in order to estimate its effectiveness and the best arrangements to use. Combination according to method (2) was used and it was found that diversity reception was more effective for pure C.W. than it was for I.C.W. This might be expected because when I.C.W. is used the transmission covers a band of frequencies and this in itself tends to overcome the type of fading which diversity is designed to combat.

An extended range of tests have been carried out to determine the best spacing to use, and also the minimum spacing which will give useful results. It appears that the spacing required (expressed in wavelengths) falls as the wavelength is increased —a very fortunate result—since it means more or less equal actual spacings for the different wavelengths.

The results obtained on the New York circuits may be summarised thus :

Wavelength (Metres).	Best Spacing.	Minimum Useful Spacing.
* 15·89	20λ	7λ
22	10λ (6λ good)	4λ
40·5	7λ to 10λ (varies)	4λ (good)

Spacings are between centres of each array.

As a result of the observations, it was concluded that if space is allowed in laying out a station for a 30λ spacing on 14·5 metres, there will be ample room for suitable spacings on all other wavelengths comprised within the short wave band. Experience suggests that much smaller spacings can produce some improvement.

The diversity principle has been found of the utmost value

* In this case each array was a Franklin Beam (Uniform aerial type), $3{\cdot}4\lambda$ wide.

when signals are strong (in average value), but suffering from severe rapid fading. Under these conditions even two single aerials suitably spaced may be nearly as good as a single large array. On the other hand, when signals are weak, the much larger " pick-up " of the array is most valuable, enabling traffic to be carried on, whilst a single aerial gives a barely audible signal at the receiver output. It would appear, therefore, that the soundest and most economical receiving system should consist of medium-sized arrays suitably spaced. The use of a very large receiving array does not appear desirable, since its polar diagram is very sharp, whilst the direction of incoming signals is liable to change under some conditions, and also the phase of the incoming signal may not be the same over the whole width of the array.

A typical combining circuit will be described in the next chapter.

TYPICAL RECEIVERS

HAVING discussed the general problems of reception and some of the circuits used, it remains to describe typical receivers. A typical "straight" receiver has already been described in the previous chapter, whilst in Chapter XVIII will be found a description of a receiver for F.–M. communication. Receiver design for some very short-wave applications can hardly be said to be sufficiently settled to allow of "typical" examples being given.

In view of this, and the space limitations, only two receivers will be described in this chapter—a communications receiver and the R.F. portion of a television receiver.

The Marconi Receiver, CR150

This receiver will now be described, as representing one of the more elaborate communication receivers. The frequency

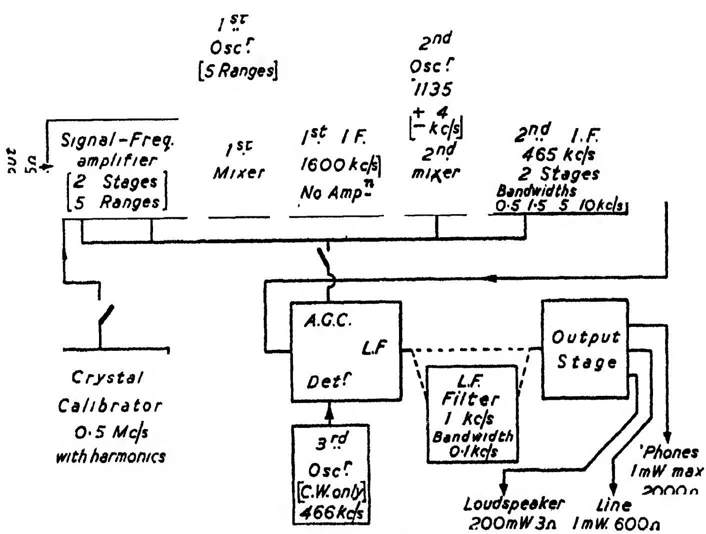


FIG. 353. Block Diagram of Marconi Receiver CR150.

band covered is from 2 to 60 Mc/s, in five ranges. By careful choice of valve types and circuit design, together with the provision of sufficient amplification at the signal frequency, the inherent receiver noise is, at all frequencies below 30 Mc/s, reduced to the theoretical limit set by the thermal noise in the input. From 30 to 60 Mc/s the first valve noise is the limiting factor.

The block diagram of Fig. 353 shows that there are two Signal-Frequency Amplifiers, followed by a mixer with separate oscillator. The first I.F. is 1,600 kc/s, there being no amplification at this frequency but the coupled tuned circuits assist in reducing second-channel interference.

A triode-hexode produces the second I.F. of 465 kc/s, for which there are two amplifying stages. The main adjacent-channel selectivity is provided here, crystal-filter circuits being used for the narrower bandwidths. A double-diode triode rectifies and amplifies the signal and provides bias for automatic gain control. A Third Oscillator can supply a frequency differing from the second I.F. by about 1 kc/s, for the reception of C.W. by the heterodyne method. If interference

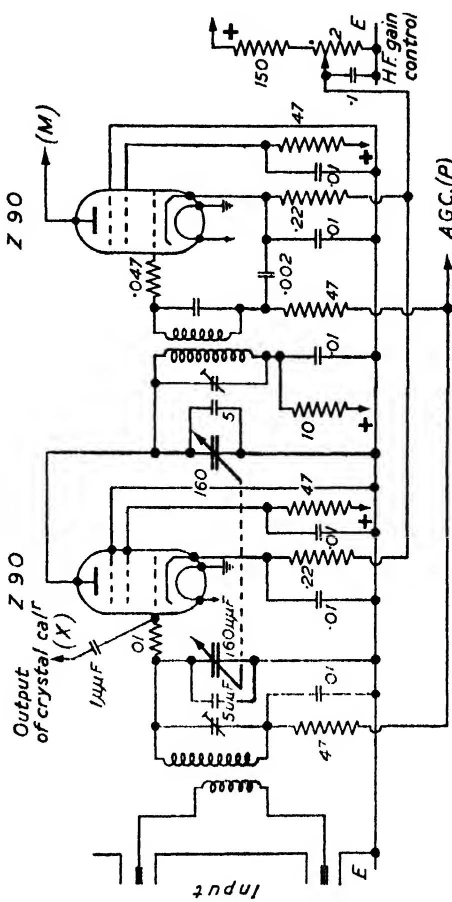


FIG. 354. Signal-frequency Amplifier of CR150.
Resistances in kΩ, Capacitances in μF unless otherwise marked.

is severe and both the transmitter and the receiver First Oscillator are sufficiently steady, a L.F. filter can then be switched in, having a bandwidth of 0·1 kc/s at a mid-band frequency of 1 kc/s.

An output stage can supply a line, loudspeaker and earphones simultaneously, and removing one output does not appreciably change the volume of the others.

We will now consider further some of the more interesting features of the receiver. A simplified diagram of the Signal-Frequency Amplifier is shown in Fig. 354, no waveband

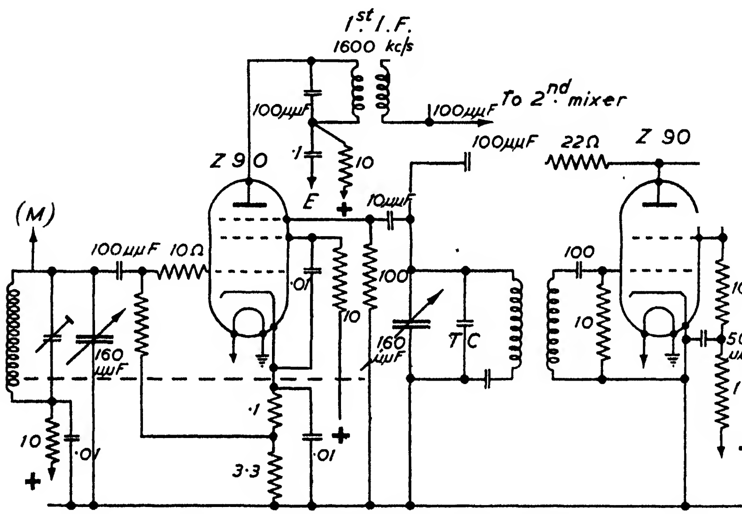


FIG. 355. First Oscillator and Mixer of CR150.
Resistances in kΩ, Capacitances in μF unless otherwise marked.

switching being shown. There are actually five sets of coils, with their own inductance trimmers and condenser trimmers.

It will be seen that two concentric sockets are used for the input. If this is balanced, one line will be connected to each, whilst if an unbalanced input, such as a concentric line, is being used, this will be connected to one socket and the other will have its central connection shorted to its outer. The input should have an impedance of 75 ohms.

The amplifier circuits are designed to produce, as far as possible, the same gain over the whole of each waveband. The gain of a pentode, tuned-anode circuit is given by $g_m \omega L Q$.

Since Q will not change much over the range of frequency covered by the variable condenser, it follows that the gain will be considerably greater for the highest frequency in each waveband than for the lowest. In this receiver coupled circuits are used, the primary being tuned by the ganged condenser but the secondary being tuned permanently to a frequency lower than the lowest frequency of the waveband. The secondary circuit therefore responds better to the lower frequencies and can be made to compensate approximately for the falling off in gain of the tuned primary circuit.

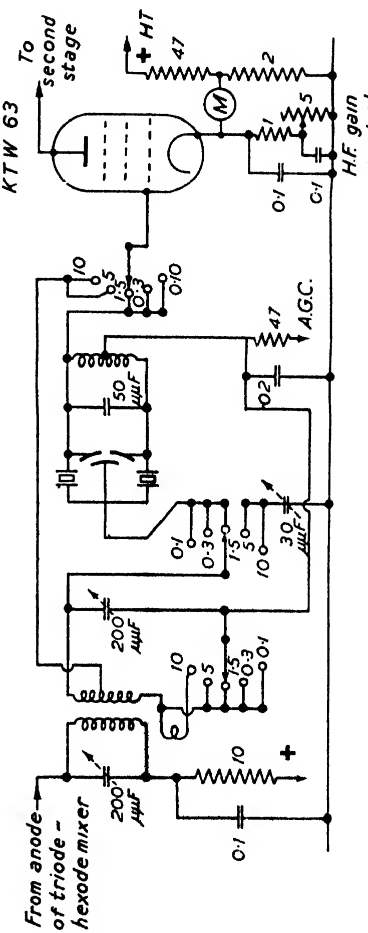


FIG. 356. Second I.F. of CR150.
Resistances in kΩ, Capacitances in μF unless otherwise marked.

Automatic gain control is applied to both valves and there is a manual control of gain, obtained by varying the cathode potential, and hence the grid bias. Special pentodes for high frequencies—Z90's—are used in the amplifier. These are similar to the EF50's, some particulars of which are given on p. 595.

Fig. 355 shows the First Oscillator and Mixer, the five wave-bands not being shown. It will be seen that the Z90 valve has been connected as a triode and a tuned-anode circuit with separate reaction coil is used. Considerable attention has been given to this oscillator, in order to make it as stable in frequency as possible. The section of the four-gang condenser, which belongs to the oscillator circuit, has a wider

spacing between its plates and they are of thicker material, to provide increased rigidity. The coils are wound on ceramic formers and a temperature-compensating condenser (T.C.) is fitted.

It will be seen that the oscillator output is fed into the mixer by way of the suppressor grid, which has a negative bias with respect to the cathode.

A diagram of the Second Mixer is not given, as it is very similar to that shown in Fig. 345. This is a change between two fixed frequencies, of course, but provision is made for about a 4 kc/s variation of the oscillator frequency by a front-of-panel control. This provides fine tuning over a range of 4 kc/s, whatever the frequency band in use.

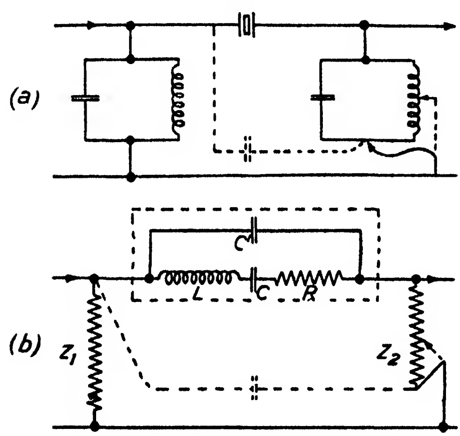


FIG. 357. Typical Crystal Filter.

Turning now to the second I.F. of 465 kc/s, the diagram of Fig. 356 shows the connections between the mixer and the first amplifying valve, for various positions of the Selectivity Switch. It will be seen that the 10 kc/s bandwidth is obtained by increasing the coupling between the circuits, whilst the narrower bandwidths employ a crystal filter.

The circuits between the two amplifying valves are identical, except that the crystal filter in this stage only comes into circuit on the 0·3 and 0·1 kc/s positions, the connections for the 1·5 kc/s position being the same as for 10 and 5 kc/s. The connections throughout the I.F. stages are the same for 0·3 and 0·1 kc/s but the same knob puts the L.F. filter into circuit on the 0·1 kc/s position. The coupling between the second I.F. valve and the diode detector consists of coupled circuits which are not varied in bandwidth.

We will now discuss, in simple terms, the performance of the crystal filter. Let us first consider the simple arrangement

of Fig. 357*a* with its equivalent circuit, Fig. 357*b*. We have seen previously that the crystal slice can be represented as a series circuit in which the L/C ratio and the Q are very high. The capacitance between the holder electrodes is C'. The resonance curve of the crystal will be so sharp compared with that of the tuned circuits, that we can regard them as in

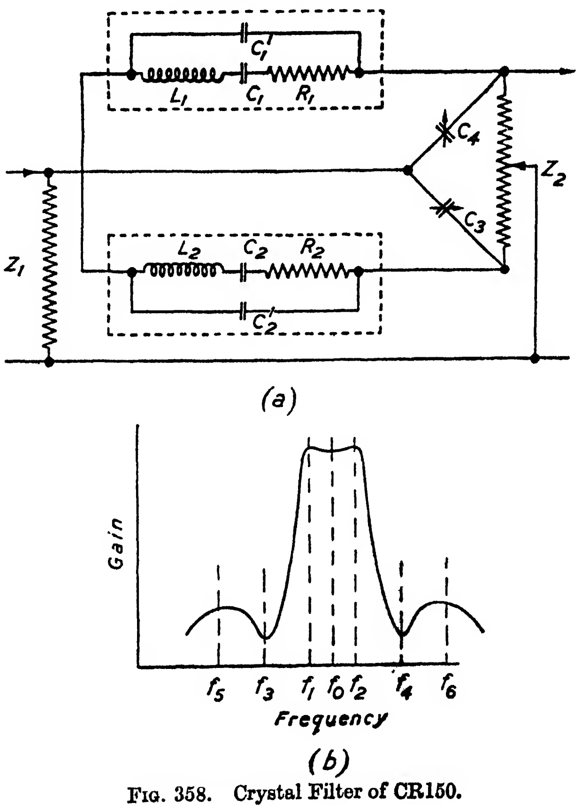


FIG. 358. Crystal Filter of CR150.

parallel resonance over the whole band of frequencies in which we are interested and, therefore, they are represented by their effective resistances $(\omega_o^2L^2/R)$, Z_1 and Z_2.

If a voltage at the *series* resonance of the crystal is applied across Z_1, it will be seen that this will pass easily through to Z_2 and appear at the output. A very slightly different frequency will, however, meet with a large reactance.

At a somewhat higher frequency the crystal in its holder will come into *parallel* resonance, with C' as the tuning capacitance. The impedance between Z_1 and Z_2 will now be an exceedingly high resistance and this frequency will be very greatly attenuated at the output.

The frequency at which this rejecting action takes place can be varied by partly neutralising C' with the balancing condenser, shown dotted in Fig. 357*b*. This balancing condenser can either be pre-set to give a definite bandwidth, or may be made a front-of-panel control, so that an interfering

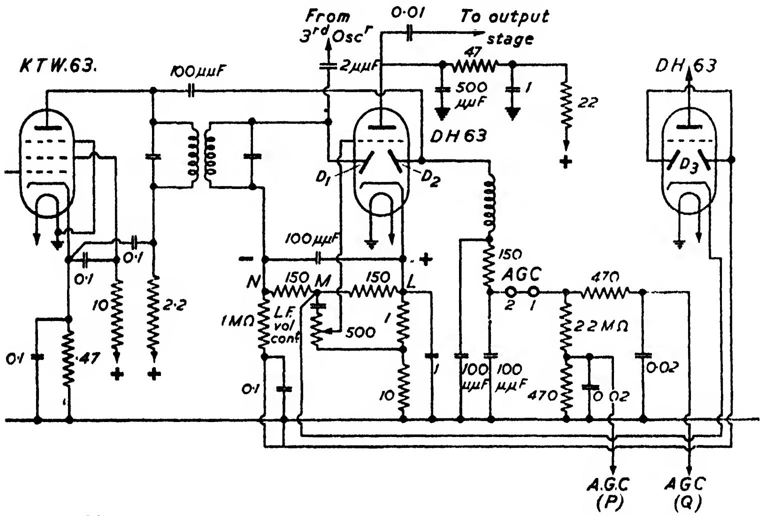


FIG. 359. Detector and Automatic Gain Control of CR150.
Resistances in kΩ, Capacitances in μF unless otherwise marked.

frequency can be tuned out, but in this case the adjustment is very critical.

Turning now to the actual circuit used in the CR150, it will be seen that this can be represented by the equivalent circuit of Fig. 358*a*. The two crystals are arranged to have slightly different frequencies of series resonance and these are the frequencies f_1 and f_2 of Fig. 358*b*. By varying the differential condenser, C_3, C_4, the position of the rejection frequencies, f_3 and f_4, can be varied, but if they are brought very near together, in order to give the curve very steep sides, then the response at f_5 and f_6 becomes greater. The usual working

condition is that which makes the response at f_5 and f_6 20 db less than that at f_0. The adjustment of C_3 C_4 is, in this case, pre-set and the alignment of the I.F. circuits can only be satisfactorily carried out with response-curve tracing apparatus embodying a cathode-ray oscillograph.

It is assumed that the reader is familiar with the diode detector of the "peak" type and that no detailed comment upon this part of the circuit is necessary.

The Automatic Gain Control circuit (Fig. 359) will be considered further, however, as being typical of A.G.C. systems. D_2 is used to produce a D.C. voltage dependent upon the signal at this point. This voltage is then applied as additional bias to the signal-frequency amplifiers and I.F. amplifiers. If a large control is required from a single valve, this should be of the "variable mu" type, in which the mutual conductance decreases gradually as the negative grid bias is increased. Hence, if the voltage to the diode rises, the gain of the preceding amplifiers falls so that the increase of voltage at the diode (and, therefore, at the output) is not so great as the increase of input signal.

The diode used for A.G.C. has a bias of about 12 V, provided by the anode current of the triode portion of the valve. It will be seen that if the I.F. voltage has a peak value less than 12 V, then no current will flow through the diode and I.F. currents will flow through the condensers to earth, no D.C. voltages being developed across the resistances.

When the I.F. voltage is greater than 12 V, however, the current will flow mainly through the diode during a portion of each positive half-cycle whilst during the negative half-cycle it will flow through the load. Since this current through the load is now unsymmetrical, it will produce a D.C. potential across the resistances, and hence an additional bias on the amplifiers.

The connection marked AGC(P) goes to the point bearing the same label in Fig. 354, whilst AGC(Q) goes to the point marked AGC in Fig. 356.

Thus for signals at the diode below a certain value, the gain of the amplifiers will remain a maximum, but above this the gain will decrease as the voltage applied to the diode increases. In this receiver the output does not increase by more than 9 db

for an increase of 60 db in the input signal above the value necessary to give a satisfactory signal/noise ratio.

If telephony is being received, then the voltage rectified by the diodes is mainly due to the carrier, and the gain of the amplifiers therefore varies according to the carrier voltage. It is desirable that the A.G.C. should have a time constant associated with it as otherwise it might start varying the gain very rapidly and thereby causing noise and distortion. In this receiver the time constant (depending upon the condensers and resistances forming the diode load) used for telephony is 0·2 secs. When receiving C.W. telegraphy, the carrier is only on, of course, during " marks " and the gain would therefore rise between each signal element unless a sufficiently long time-constant is used. In this receiver the time-constant is made 1·75 secs. for reception of hand-speed Morse. If the receiver is to be used to supply a high-speed recording unit (see page 666), this is reduced to 0·5 secs.

A noise-limiting circuit is provided, which gives a measure of protection against impulsive noise. This utilises the diode portion (D_3) of a double-diode triode valve, of which the triode portion is used in the frequency calibrator (described below).

It will be seen that the load resistance LN of the signal diode D_1 is centre-tapped. The portion LM is used to supply the L.F. output in the usual way, whilst MN normally keeps the anode of D_3 negative, so that it passes no current. The anode voltage of D_3 does not follow the modulation, owing to the long time-constant of the 1 MΩ and 0·1 μF.

Suppose that an impulse having a peak amplitude greater than that of the signal carrier now arrives. This will cause M (the cathode of D_3) to become very negative, but the anode of D_3 does not change in potential and hence current flows in D_3. For a sudden change, the 0·1 μF condenser is nearly equivalent to a short-circuit and therefore the low resistance of the conducting diode D_3 is placed across the L.F. output, practically shorting it.

The crystal calibrator (see Fig. 360) comprises an A.T. cut, 0·5 Mc/s crystal, in a circuit which gives an output rich in harmonics. When desired, this output is fed into the signal-frequency amplifier and, with the Third Oscillator on, beats

can be heard at 0·5 Mc/s intervals, up to 30 Mc/s. Thus the receiver settings can be checked over most of its range.

Considering the circuit of Fig. 360 in conjunction with the equivalent circuit of a quartz crystal given on page 586, we see that at the series resonance of the crystal the circuit is practically that of a Colpitts oscillator, whereas at other frequencies a high reactance would be placed in the grid circuit. The arrangement therefore oscillates at the crystal frequency. The object of the coil L is to increase the harmonic output. It will be seen that the output is taken across L and the voltage output will therefore be increased at the higher harmonics, since the reactance of L rises. L will, naturally, have stray capacitance

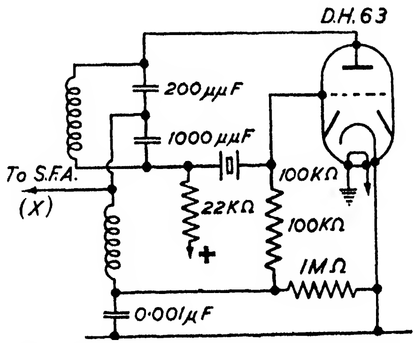


Fig. 360. Crystal Calibrator of CR150.

associated with it and, when parallel resonance is reached, the output will be good, but for crystal harmonics higher than this will fall rapidly.

The satisfactory and convenient operation of a receiver is very dependent upon a good design for the tuning control. In the CR150, a linear scale, calibrated in frequency, is used, together with a "logging scale." The frequency scale is seen in the long window at the top of the receiver front panel (Fig. 361). The scales for the five ranges are arranged around a cylinder and the band-change switch also turns the roller, so that only the scale in use can be seen.

The outer part of the tuning knob turns the four-gang condenser spindle through 25/1 gearing, spring-loaded to prevent backlash. The inner part of the knob provides a further 6/1 gearing, giving 150/1 altogether, when required.

The lower scale in the circular window is driven direct from the outer part of the knob and each half-circumference of it carries 10 main divisions. Hence, when the condenser spindle turns through one-half of a revolution (minimum to maximum capacitance) the lower scale will go round $12\frac{1}{2}$ times, or $12\frac{1}{2} \times$ 20 main divisions. As each main division has five sub-divisions, this gives a total of 1,250 divisions on a scale 18 feet long. When once a station has been found on this logging scale, the receiver can be re-tuned to it with great accuracy.

The function of the upper scale in the circular window is to

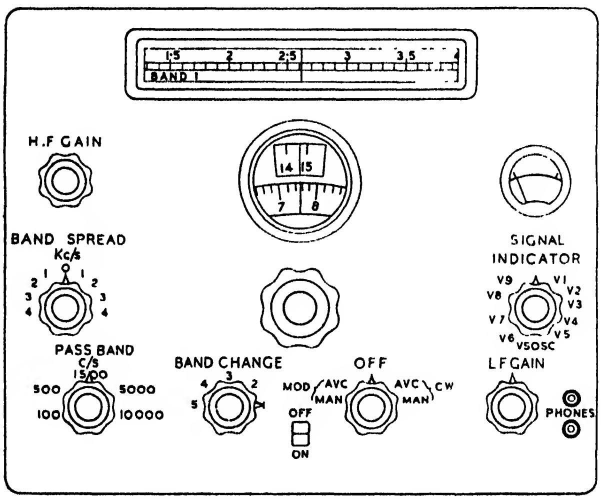


FIG. 361. Front Panel of CR150.

count the revolutions of the lower scale, so that it is graduated from 0 to 25 and each division corresponds to half a revolution (10 divisions) of the lower scale. The disc carrying the upper scale also drives the hairline across the frequency scale, by means of a cord.

From the description given, it should be possible to understand the function of the various controls. The Operational Switch, when on C.W., switches on the Third Oscillator and allows a choice between manual control of gain or A.G.C. with long time-constant. Similarly, on the modulation side there is a choice between manual control of gain or A.G.C. with short time-constant. The off position switches off anode voltage

only and therefore leaves the receiver warmed up, so that the frequency does not drift when it is switched on again.

It will be seen that a meter is provided which reads the feed currents of all the valves. Shunts are provided so that for all positions of the meter switch the readings fall between the same limits if the valves are in good order.

This meter is also arranged to give an indication of signal strength, the circuit for this being included in Fig. 356.

When no signal is being received, the tapping on the potentiometer provides the same p.d. as that due to the anode current through the bias resistor. When a signal is large enough to cause the A.G.C. to function, then the anode current will decrease and hence a current will flow through the meter. The relation between grid voltage and anode current is roughly logarithmic in these variable-mu valves and hence the meter records signal strength according to an approximate decibel scale. The H.F. Gain Control is used near its maximum and varied slightly to give zero reading on the instrument, in the absence of a signal.

Diversity Reception with the CR150 Receiver

The general principle of diversity reception has been touched upon in the previous chapter. As the CR150 has been greatly used for this type of reception, the suitable methods will be studied by reference to it.

The standard equipment utilises three CR150 receivers, mounted in a cabinet, together with the combining circuits and recording units for high-speed telegraphy.

It is necessary that the three receivers should use the same First Oscillator, so that the I.F.'s are necessarily the same. The standard CR150 has a plug socket for feeding in the output of a common oscillator, instead of using the one in the receiver.

The design of the common oscillator is very similar to that in the original receiver and is followed by a buffer valve and power amplifier. Three concentric cables then connect to the receivers.

It is also necessary that the A.G.C.'s should be combined. If this were not done, then the receivers which were receiving a small signal at any moment would have the gain raised to a high value and would contribute a lot of noise to the combined

output. The individual receivers have terminals A.G.C.1, 2 (see Fig. 359) by which their A.G.C. circuits can be broken and brought to a combining unit.

From Fig. 362 it will be seen that the current flowing through the 470kΩ and 2·2MΩ resistances will be the sum of the currents provided by the diodes in the three receivers. Hence, if the signal on one receiver is predominating at any moment, this will mainly determine the negative voltage applied to the grid of the double-diode triode. The triode portion of this is a cathode-follower, the lower end of the cathode resistance being at −120 V with respect to earth. In the absence of a signal, the anode current would be sufficient to put the cathode at a

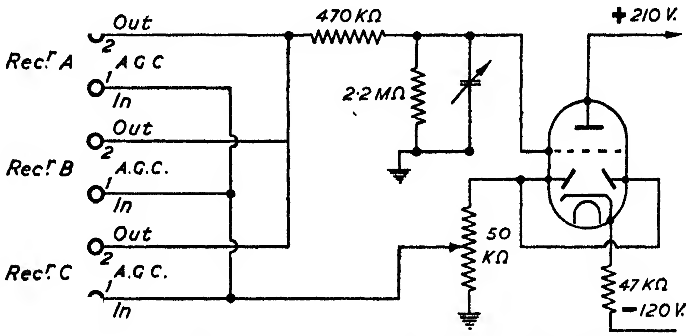


FIG. 362. Arrangement of A.G.C. for Diversity Reception.

higher potential than the diodes and hence the 50kΩ potentiometer would carry no current and no bias voltage would be put back into the receivers. If a signal makes the grid go more negative, however, the anode current is reduced, the cathode becomes less positive, current flows through the potentiometer and bias is applied to the receivers.

The time constant of the A.G.C. circuit can be varied by switching in various values of capacitance across the 2·2MΩ resistance. The time constants in the individual receivers is set to the smallest value, so that the time constant is mainly determined by the adjustment in the combining unit. By employing the cathode follower in this way, the output resistance can be of a low value (50kΩ) and hence stray capacitances in the long leads to the individual receivers do not affect the time constants.

In the case of telephony, the V.F. output of each receiver is passed through a volume control and then applied to the grid of a valve. The anodes of all three valves are strapped together and connected to an output transformer, and so to line.

If high-speed telegraphy is to be recorded, or sent over a line to a Central Office, the Recording Circuits described in Chapter XVII are used. Additional double-diode rectifiers are provided for the two extra receivers and the three rectified outputs are combined in a common load resistance in the Recording Unit.

A Television Receiver—Pye B16T

Owing to limitations of space, this book does not deal with television but it is thought that a very brief description of the R.F. circuits only, of a post-war television receiver may be of interest. The special features of a television receiver will be due to the high carrier frequency and high modulation frequency —a bandwidth of at least 4 Mc/s being required.

As has already been mentioned, the British television receiver has only to receive on two fixed frequencies (vision 45 Mc/s and sound 41·5 Mc/s),* and hence a "straight" receiver is a practical design and the Pye receiver is of this type.

Following the usual practice, the same $\lambda/2$ aerial is used for vision and sound, connected to the receiver through a concentric feeder. The first two amplifying stages are also common and then both vision and sound have each two further stages to themselves. The valves in all the R.F. stages are EF50 pentodes employed because of their high g_m (6·5mA/V) high input impedance (about 5000Ω at 50 Mc/s) and low capacitances ($C_{ag} < 0{\cdot}003\ \mu\mu F$).

If a reasonable amplification is to be obtained with such a large bandwidth, the tuning capacitance will have to be kept to a minimum and in this receiver no condensers are connected across the tuning coils, the largest part of the tuning capacitance being that of the valves. The resonant frequency is adjusted by moving iron-dust cores in the coils.

Since the frequency separation between vision and sound is only 3·5 Mc/s and the maximum vision modulation-frequency is 3 Mc/s, the circuits following the common stages have to be very carefully designed, if the full range of vision modulation

* Frequencies of Alexandra Palace transmissions.

is to be received without interference from the sound signal. Experience has shown that at least 40 db suppression of the sound signal is necessary, to avoid effects on the picture.

The necessary "steep-sided" response curve is partly

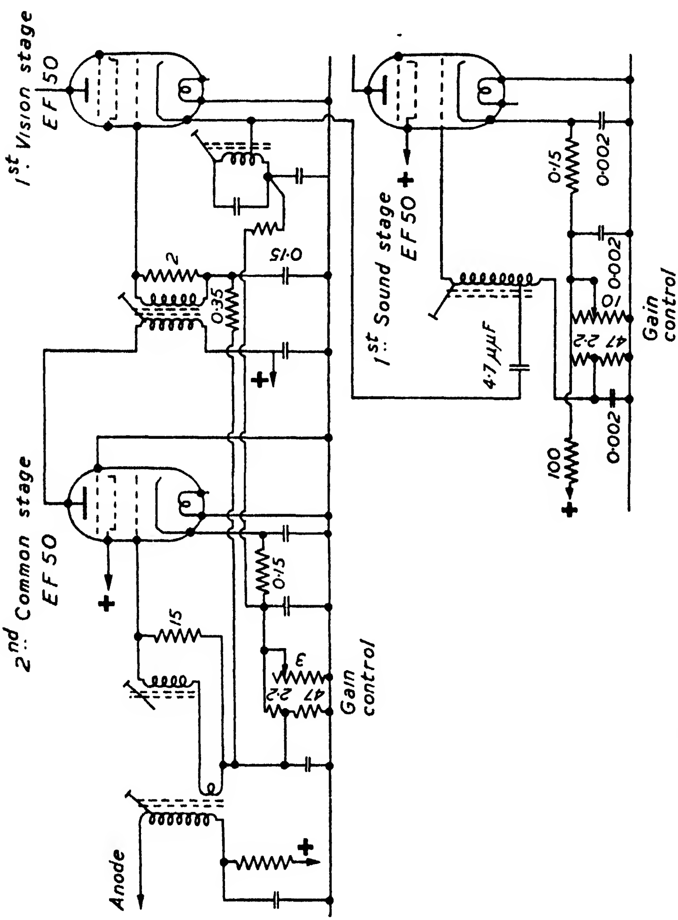


FIG. 363. Circuit Diagram, Part of Pye B16T Television Receiver.
Resistances in kΩ, Capacitances in μF unless otherwise marked.

obtained by circuits in the cathodes, tuned to the sound channel, Fig. 363 showing the circuits between the second common stage and the separate stages.

A commercial television receiver must provide for a large manual adjustment of gain, because receivers in different locations will have such different input signal strengths. The

circuits for providing this gain control require careful planning in the case of an U.S.W. "straight" receiver—in fact, the design adopted will largely depend upon the gain control arrangements. The usual variable-mu valve has not a high enough average g_m to be used in these wide-band amplifiers, and so the best use has to be made of a straight pentode, such as the EF50, but, fortunately, there are several stages to which the control can be applied. In this receiver there are three separately adjustable controls.

When the grid bias on a valve is increased, the grid/cathode capacitance falls and, since this is the main tuning capacitance, the frequency-response curve will change. The input imped-

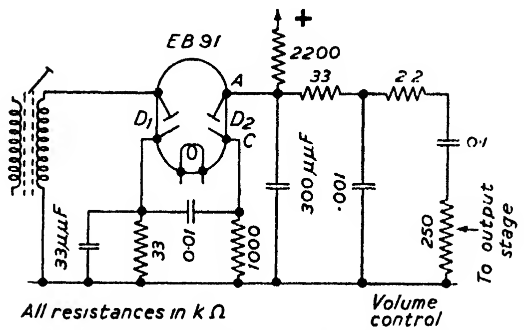


FIG. 364. Detector Circuit of Sound Channel in Pye B16T Receiver.
Resistances in kΩ, Capacitances in μF unless otherwise marked.

ance rises because of the reduction of g_m (as can be seen from page 398). The anode current falls, of course, and if the gain control is being applied to a number of valves, the output voltage of the power pack will usually rise considerably. These difficulties are largely overcome by varying the voltage of the suppressor-grid as well as the control grid. This is done by inserting the variable resistance in the cathode lead (Fig. 363). By adopting a correct ratio for the voltages applied to the two grids, the change in capacitance is greatly reduced and the total current drawn by the valve does not vary so much, because the screen current increases.

Diode detection is employed for both vision and sound and in each case the second diode in the valve envelope is used in a circuit for reducing the effects of interfering impulses. The detection stage for the sound channel is shown in Fig. 364,

and it will be seen that the output of D_1 has to pass through D_2 in order to reach the output stage. D_2 is normally conducting because it has positive bias derived from the anode supply. It will be seen that signals and interference are tending to make C positive.

Leaving out of account, for the moment, the effect of any time-constants in the circuit, it will be seen that normal signals would leave D_2 still conducting, but a signal of large peak value would make C more positive than A and D_2 would cease to conduct, for the duration of this signal.

The rate of change of potential at A is limited by the time constant of the 2·2 MΩ resistance and 300 $\mu\mu F$ condenser and these values are such that the potential can follow the maximum amplitude and frequency of the wanted signal, but the impulses grow more rapidly and the potential is unable to follow.

The arrival of an impulse means, therefore, that C rises rapidly but A rises less rapidly, and hence an impulse of smaller value than the signal may be cut off by D_2.

Most impulsive interference is of such short duration that the break in the sound output is unnoticed, whereas a large impulse in the output stage may shock the loudspeaker into vibration as well as cause distortion by overloading. The time constants of the D_1 circuit are made smaller than usual, so as to preserve the " peakiness " of the impulse.

CHAPTER XVI

WIRELESS TELEPHONE CIRCUITS

In recent years the telephone networks of the principal countries have been linked up by international circuits. Where long distances across the ocean have to be bridged, wireless is exclusively used, and it is these channels that we propose to discuss in the present chapter.

The complex special equipment that is necessary in such a circuit is often a matter of some surprise to the telegraph engineer. Accustomed as he is to duplex or even multiplex working, he finds it strange in reverting to telephony, which is virtually simplex, that before his existing transmitters, receivers and array systems can be utilised he has to install a very complex and costly terminal equipment.

It may not be out of place therefore to review the problem generally, casting back to the simplest form of telephone channel. In the early days before line amplifiers, or repeaters as they are called, a telephone channel consisted of a 2-wire line, at each end of which was a "pair" consisting of a carbon-microphone and polarised ear-piece, the circuit in its simplest

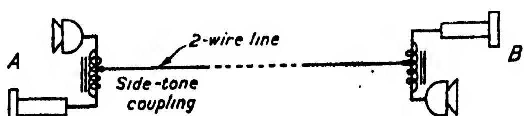


Fig. 365. Simple Telephone Circuit.

form being shown schematically in Fig. 365. When A., for instance, speaks, the current can travel not only along the line to the receiver at B., but a small portion of A.'s speech is diverted through a transformer into his own ear-piece. This by-passed speech is known as side-tone, and the side-tone adjustment is very valuable. Its prime function is to give the speaker confidence, as the average individual hearing his own speech assumes it is being correctly transmitted. In addition side-tone tends to control the level of speech input of

the average individual who is neither deaf nor super-sensitive of hearing. And it can be used to level up a transmitter-receiver pair. Thus by giving a bad microphone a lower side-tone coupling the speaker will automatically raise his own speech level until the side-tone appears normal.

If the equipment at the far end does not correctly terminate the line, some of A.'s speech will be reflected back to his own receiver, but with a short line such as we are discussing any

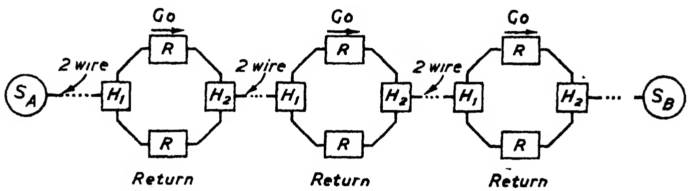


FIG. 366. Two-wire Line with Repeaters.

reflected speech will merely add to the side-tone amplitude, since the phase difference of the original speech and its reflection in a short line is negligible.

With such a circuit, since speech currents are free to travel in either direction at any time, it is a duplex system. In fact, A. and B. could properly sing a duet for each to hear. But it is important to realise that the conversational nature of telephonic communication has by established usage made its character most definitely one of question and reply, and these requirements can be adequately fulfilled by a simplex circuit. For, quite obviously, no useful purpose is served by allowing both persons to speak at once.

Since the only amplifier in the circuit is the microphone, early telephone systems such as we have described were limited in range by the line attenuation and noise level. Generally speaking this meant a maximum of some 25 miles and networks therefore grew up around the large cities.

Although some lines of specially low attenuation were used for long-distance telephony, this really only came into its own after the introduction of the valve amplifier.

Since a valve amplifier can only work in one direction, two systems were evolved, which are shown in outline in Figs. 366 and 367.

In the 2-wire system, " repeaters " would be inserted at

intervals along the line, each repeater including arrangements H_1, H_2 (to be discussed later) for separating the speech currents passing in opposite directions and passing them through amplifiers " facing " in opposite directions.

In the 4-wire system, different pairs of lines are used for speech in the two directions and the repeaters therefore become

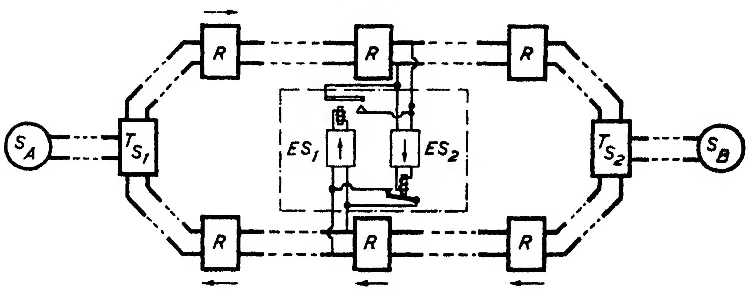


FIG. 367. Four-wire Line with Repeaters.

straightforward amplifiers. Apparatus TS_1, TS_2, for separating out the " go " and " return " speech, is then only necessary at the ends of the 4-wire circuit.

The introduction of line repeaters leads to difficulties due to self-oscillation (termed " singing " by the telephone engineer), whilst the extension of the line to great distances introduces two inherent troubles, delay-time and echo. We propose to discuss these in turn.

Singing

Any amplifier can be made to self-oscillate by sufficient feed-back, and it was shown on page 416 that when the gain round the loop formed by amplifier and feed-back exceeds the loss, self-oscillation will occur.

It will be seen that in all the telephone circuits which we have discussed, even the simplest, there is a loop circuit containing an amplifier. In the case of the circuit of Fig. 365 loops are formed at each end by the local subscriber's circuit. This loop contains the microphone (which is an amplifier), side-tone coupling to the receiver (largely neutralised in modern instruments) and an acoustical link between earpiece and mouthpiece. In the modern hand-set the receiver and

microphone are so fixed that the attenuation exceeds the gain but with the old "candlestick" instrument, singing could be easily obtained by placing the earpiece within the mouthpiece. We hasten to add that this practice is most unpopular with the G.P.O.!

In the case of a 2-wire repeater, the loop attenuation introduced by H_1 and H_2 must exceed the gain of RR, and this will require very carefully balanced circuits at each repeater, as will be clearer after the differential transformer has been studied. For this reason, amongst others, the 2-wire repeater is practically obsolete.

With the 4-wire circuit there is a loop around the complete circuit, even though it reach from New York to Mexico, and for this loop all the repeaters are facing the same way. Thus in all such loops, whether it be the local loop of a 2-wire or the extended one of a 4-wire system, we must insert into the loop, at the junction of the incoming and outgoing lines, a network which will put as large an attenuation into the loop as possible but designed so as to put the least possible attenuation into the correct paths of the "go" and "return" currents.

Differential Transformers

The equipment used at H_1, H_2, TS_1, TS_2, to fulfil the above requirements, is a differential transformer, often known as a "hybrid coil." One type is shown in Fig. 368, connected between a local subscriber's line and "go" and "return" circuit amplifiers. It will be seen that the transformer is coupled, not only to the subscriber's line and to the "go" and "return" lines, but also to a balancing network, the components of which can be adjusted to simulate the input impedance of the 2-wire circuit.

Assume that in the transformer of Fig. 368 all the coils have the same number of turns and are wound in the same direction. Suppose that speech is coming in on the "return" line. Then the currents flowing in PP' and QQ' induce E.M.F.'s in MM' and OO' and, if the balancing network has exactly the same impedance as the 2-wire line, the currents indicated by the dotted arrows are equal. It will be seen that equal but opposite E.M.F.'s are therefore induced in JJ' and KK', so that no currents flow in the "go" line. Half the energy

coming in on the " return " line is evidently wasted in the balancing network, so that an attenuation of 3 db has been inserted between the " return " line and the 2-wire line.

Now suppose the local subscriber speaks into the two-wire line. E.M.F.'s are induced in JJ' and PP' and, assuming that the input impedances of " go " and " return " lines is the same, equal currents will flow in the " go " and " return " lines. The energy flowing outward along the " return " line will be dissipated in the termination of the first amplifier (" facing the

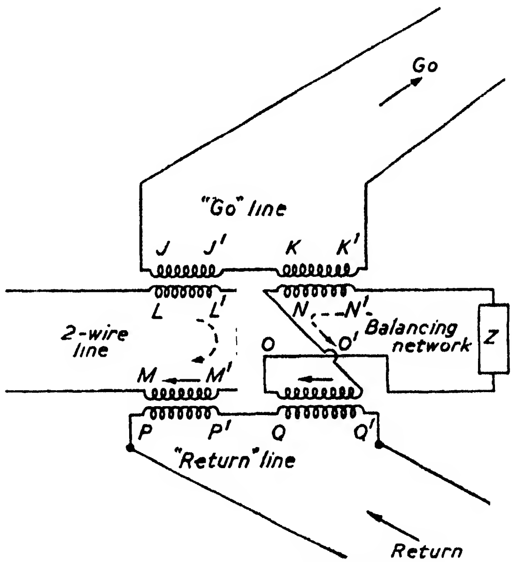


FIG. 368. Differential Transformer.

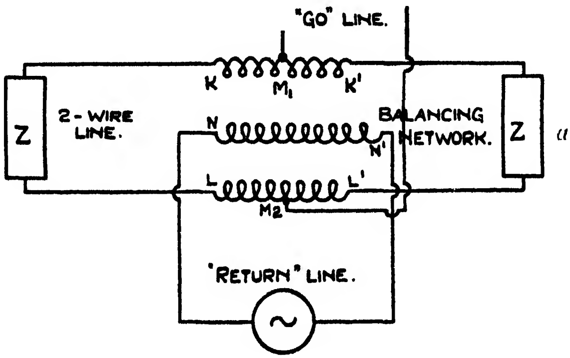
other way ") which it meets. Half the energy from the 2-wire line is therefore wasted, as before.

An alternative form of differential transformer is shown in Fig. 369*a*. Again considering an input from the " return " line, an E.M.F. is induced in LL' and KK', and this circulates a current through the balancing network, and the 2-wire line, in series. Since the combined impedance of M_2L', the balancing network, and $K'M_1$ is equal to the combined impedance of M_1K, the 2-wire line, and LM_2 and equal E.M.F.'s are induced in each half winding, there will be no P.D. between M_1 and

M_2. If necessary, reference to a similar circuit (Fig. 369*b*) will make the matter clear.

The first type of transformer is somewhat more costly in material than the second but each winding can be manufactured to wider tolerance and yet a good balance obtained by suitable pairing up of the coils after test.

It might be thought that this provides all that is required for satisfactory operation but difficulties would arise in practice, owing to imperfect balance. An actual differential transformer



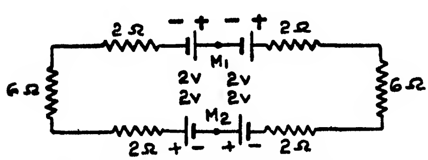


FIG. 369. Differential Transformer.

will usually only insert about 80 db into the loop circuit if it is connected to perfectly matched impedances, but when it is in use at the end of a 4-wire circuit the 2-wire extension will undergo wide variations of impedance as different subscribers are connected. Consequently, the loop attenuation will fall at times to a much lower figure.

However, it is found possible to work long 4-wire lines with zero loss and still to have an adequate margin of safety against singing, that is, the output to the 2-wire extension at one end is at the same level as that being received from the 2-wire extension at the other.

Transmission Time

The transmission time, that is, the time taken for a given portion of the speech wave to travel from one subscriber to another, should be as small as possible. When a subscriber has finished talking, at least twice the transmission time must elapse before he can receive a reply. During this period of uncertainty both persons, if not used to the conditions, are likely to try and talk simultaneously, causing confusion and, on circuits fitted with suppressors (to be discussed later), a lock-up of the system.

On telephone lines the velocity of transmission ranges from 16,000 km/sec. on heavily-loaded lines (that is, lines with inductance added to reduce distortion) to nearly 300,000 km/sec. on more modern, unloaded wide-band carrier circuits. It is only on very long lines that transmission-time will be troublesome. In these days telephony is world-wide but for long distances over sea, wireless is always used, for which the velocity is, of course, 300,000 km/sec. It is important, however, that on wireless telephony circuits the transmission time should not be increased by the apparatus at the terminals, especially in view of the fact that the wireless link will often be extended over long trunk lines.

Echo

We mentioned that if the line apparatus is incorrectly terminated, reflection of speech will occur, but with a short line such reflected speech merely merged with the side-tone. With a long 4-wire circuit, however, troublesome echoes can travel back from the distant junction with the 2-wire extension.

Further, even if the terminal hybrids are perfectly balanced, (should the 2-wire circuit to S_A, for example, not be perfectly terminated) reflection of speech originating at S_B will occur and when this reflected wave reaches the hybrid coil at TS_1 it will be passed back to TS_2 and eventually reach S_B as an echo of his own speech, since even a perfect hybrid coil cannot distinguish between S_B's speech reflected from S_A and S_A's speech, any echo of S_B's speech being approximately twice the transmission time after hearing the original speech.

These difficulties are overcome in long 4-wire line circuits

by working with gains which keep the circuit below the "singing" point, and by the use of "echo suppressors" in which the speech currents from one subscriber render transmission in the reverse direction impossible while they last. The channel is, therefore, no longer truly duplex, but as we have observed previously, the nature of telephonic conversation makes a duplex system unnecessary. It only makes "breaking in" more difficult.

One arrangement of an Echo Suppressor near the middle of a long 4-wire line circuit is shown schematically in Fig. 367. ES_1 and ES_2 are amplifiers energising relays which, on closing, short circuit the lines. When neither subscriber is speaking both lines are complete, but should subscriber S_1 speak, for example, a portion of his speech wave, on reaching the repeater station half-way along the line, is amplified by ES_2 and caused to close the contact short circuiting the "return" line. It will be seen that the relay need not be particularly rapid in its operation, all that is necessary is that it shall close before the echo reaches half-way back along the line.

It will be evident that with any circuit having echo suppressors it is still not possible for the gains of the repeaters to be adjusted to equal or exceed the losses in the lines plus the loss across the hybrids from "return" to "go" lines, otherwise a complete circuit having zero or negative resistance will be formed capable of self-oscillation when neither subscriber is speaking. The "singing" would be intermittent because its onset would cause the relays to close but it is obvious that the circuit would not be workable.

Should we desire to have an overall gain round our 4-wire circuit in order that the output into the distant 2-wire circuit may be considerably greater than the input (in order, perhaps, to obtain a sufficient strength at the distant end of a poor, 2-wire line) we shall need to convert our Echo Suppressor into what is known as a Singing Suppressor which always keeps one line blocked even if neither subscriber is speaking.

Such a system will overcome not only "singing" around any loop but will, of course, stop echo, but since Singing Suppressors are a development of wireless telephony we will discuss them later in the chapter.

Carrier-Current Working

Since long telephone lines are very costly, it is desirable to be able to pass more than one speech channel over a line, even if this involves more elaborate terminal equipment. This has been very successfully achieved by applying what are essentially wireless methods to line working.

The speech currents are used to modulate a carrier and the

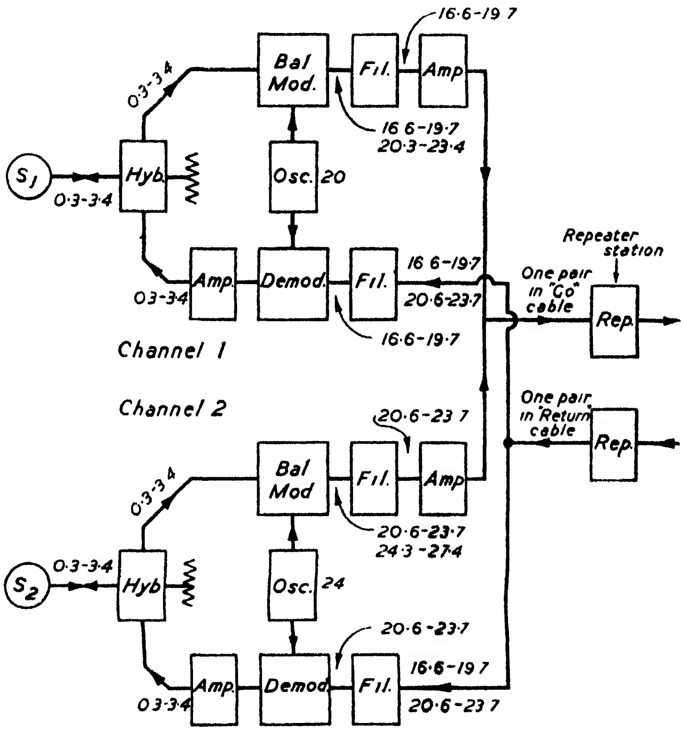


Fig. 370.

modulated carrier is applied to the line. At the far end a filter circuit accepts only the frequencies present in the modulated carrier and the original speech frequencies are recovered by detection, as in wireless. It is clearly possible for other speech channels to modulate carriers at other frequencies and the channels can be separated out by the filters at the receiving end.

In order to get as many channels into the band of frequencies which the line will pass without undue attenuation, the suppressed-carrier, single sideband system is always used.

One end of a carrier system, in which there are only two channels is shown schematically in Fig. 370, the carrier frequencies in this case being 20 and 24 kc/s. It will be seen that each channel is working on a 4-wire basis and all that has been said about such circuits still applies. From the figure it is clear that the selected single side-bands of the separate channels are automatically separated by the appropriate filters at the receiving end in conjunction with the same oscillators which thus act both to produce the transmission frequencies used, and to demodulate the received ones.

The repeaters carry all the channels and must be very free from non-linearity, or there will be cross-modulation between the channels. The successful development of carrier working has been due to the introduction of negative feed-back amplifiers, improved technique for filter design and the production of stable oscillators.

Cables are now in use containing twenty-five pairs of wires, each pair being able to carry currents up to about 150 kc/s, provided repeaters are installed about every sixteen miles. This enables twenty-four speech channels, with their carriers spaced 4 kc/s apart, to be transmitted over a single pair. Pairs of wires carrying currents in one direction will be placed in one cable, whilst a different cable will be used for the return pairs. This avoids cross-talk troubles, since the currents in the different pairs in the same cable will always be at about the same level. A more elaborate system, using concentric lines each capable of carrying 600 channels, has also been developed.

In the standard Post Office 24-channel system, each of the 24 incoming 2-wire circuits is first converted into a 4-wire circuit at the appropriate carrier frequency by the circuit shown in Fig. 370, but instead of using 24 oscillators each spaced 4 kc/s apart to obtain the 24 side-bands, the same result is obtained with 12 oscillators and the use of a second modulation (and demodulation) system for half the group. Thus the 24 speech channels are separated into two groups, 1–12, 13–24, but whereas the second group is modulated once before being passed to line as in the simple example just discussed, the first group

is modulated twice, as may be seen from Table XXIV, which sets out the modulation frequencies and side-bands.

TABLE XXIV

Speech Band = 0·3 to 3·2 kc/s

	Group 1				No. of Channel		Group 2		
	L.S.B.	2nd MOD	L.S.B.	1st MOD			1st MOD	L.S.B.	
	kc/s	kc/s	kc/s	kc/s			kc/s	kc/s	
Sent to " Go " Line and Received from "Return" Line	15·2 12·3	120 kc/s	104·8 107·7	108	1	24	108	104·8 107·7	Sent to " Go " Line and Received from "Return" Line
	19·2 16·3		100·8 103·7	104	2	23	104	100·8 103·7	
	23·2 20·3		96·8 99·7	100	3	22	100	96·8 99·7	
	27·2 24·3		92·8 95·7	96	4	21	96	92·8 95·7	
	31·2 28·3		88·8 91·7	92	5	20	92	88·8 91·7	
	35·2 32·3		84·8 87·7	88	6	19	88	84·8 87·7	
	39·2 36·3		80·8 83·7	84	7	18	84	80·8 83·7	
	43·2 40·3		76·8 79·7	80	8	17	80	76·8 79·7	
	47·2 44·3		72·8 75·7	76	9	16	76	72·8 75·7	
	51·2 48·3		68·8 71·7	72	10	15	72	68·8 71·7	
	55·2 52·3		64·8 67·7	68	11	14	68	64·8 67·7	
	59·2 56·3		60·8 63·7	64	12	13	64	60·8 63·7	

Thus we see that in channel 1, for instance, the speech frequencies 0·3–3·2 kc/s first modulate a 108 kc/s carrier. This with one side-band is suppressed and the lower side-band 104·8–107·7 kc/s selected and passed. Similarly channel 2 modulates a 104 kc/s oscillator and so on, channel 12 modulating a 64 kc/s carrier the lower side-band in each case being passed. Each of the 12 channel " go " leads from this group are then joined, as also are the " return " leads, to form a pair which is fed through a hybrid into a second 4-wire network in which a single 120 kc/s carrier is used, filters accepting only the lower side-band of the group thus formed, the total frequency range

covered when all the circuits are in operation being from 59·2–12·3 kc/s as seen in Table XXIV. This output, together with that from group 1, now to be discussed, then passes to the main " go " line.

Channels 13–24, comprising the second group, are dealt with in a similar manner except that no second group modulation is carried out and thus the frequency range of this group is the same as that produced from the first modulation of group 1, namely 107·7–60·8, as shown in Table XXIV, from which it is seen that channel 24 uses the same oscillator as channel 1, 23 as 2 and so on. The combined output from group 2 is passed to the main " go " line as has been explained above which thus carries a total frequency band from 107·7–12·3 kc/s.

It will be seen from the above that the total frequency band passing to line is exactly the same as would be obtained by using twenty-four separate oscillators spaced 4 kc/s apart except the side-bands of one group are inverted to the other. This does not matter, of course, since in the reverse demodulation process straight speech results in both cases for obvious reasons.

The advantage of this method is that both groups are initially dealt with at the higher frequencies for which crystal filters are more easily constructed. Further, the twenty-four channels are handled by fewer oscillators. It will be noted that there is no modulated carrier at 60 kc/s and this frequency is used to synchronise the oscillators at both ends.

The incoming signals which arrive on the " return " pair are dealt with in the reverse manner. Thus the two main groups are separated by filters diverting those below 60 kc/s to the common 120 kc/s demodulator of group 1, where each separate single side-band is demodulated, sorted and passes into its appropriate channel to be converted back to speech. The group above 60 kc/s passes directly to the demodulators and filters of group 2 which again divert the speech into the appropriate 2-wire line.

Besides enabling better use to be made of the line, carrier working has the advantage that the higher frequencies are transmitted along the type of cable used with a higher velocity than along the loaded cables previously necessary.

Also, since the bandwidth occupied by one channel is now only a small percentage of its mean frequency, the attenuation

and phase shift introduced by the cable is about the same for all the modulation frequencies of one channel. When the speech frequencies are transmitted directly along the cable, on the other hand, the attenuation and velocity of the low and high speech frequencies are quite different, leading to distortion unless the cable is heavily loaded. In consequence of these advantages, carrier working has become the usual method for long distances.

General Requirements of a Commercial Wireless Telephone Circuit

The ideal to be aimed at is, of course, that subscribers should not be reminded of its character or of its length by any special peculiarities or by poor performance. The output should, therefore, be at a constant and sufficiently high level, there should be an absence of noise and echo and the circuit should pass frequencies between 300 and 3,400 c/s. A narrower band, up to 3,000 c/s, may be used if circumstances so demand. Further, the wireless circuits must be so designed that the usual subscriber's instruments can be used in the normal manner. In fact, the wireless link must be considered merely as a certain form of trunk line.

The wireless circuit has two undesirable features which will make the attainment of this ideal difficult. Firstly, a signal/noise poorer than that of a trunk line, because the use of repeaters at frequent intervals is not practicable, and secondly, a transmission path which is very variable compared with a well-maintained trunk line, so that varying attenuation and, possibly, distortion are introduced.

In order to obtain a satisfactory signal/noise ratio, a highly-directional receiving system will be essential and on short waves directional transmission will also be used. Whilst a satisfactory signal/noise ratio can be obtained on a wavelength as long as 5,000 m. across the Atlantic, by the use of a single sideband suppressed carrier transmitter with an output of about 200 kW and a special form of directional receiving system, it is not economically possible to use long waves over greater distances or in directions other than away from the prevailing atmospheric centre, and short waves are invariably used.

The varying attenuation will have to be corrected for by

an efficient automatic gain control in the receiver and special methods of working are introduced to reduce distortion.

Terminal equipment of a special character will be necessary in order that the wireless circuit may be connected to an ordinary telephone circuit and considerable amplification will need to be available at the terminal points, so that losses in the land-line extensions may be made good and the wireless transmitter kept fully modulated, which is essential in order to maintain the signal/noise ratio.

Owing to the very variable attenuation of the wireless path, the ordinary echo suppressor is valueless and singing suppressors must be used.

A telephone system should be secret. When lines are used this is practically ensured by the nature of their connection but when wireless is used eavesdropping can obviously be indulged in, unless means are adopted to prevent it. Hence a "privacy" equipment, which will render speech unintelligible to all except the correct recipient, is a feature of present-day wireless telephony.

The principles of transmitters, receivers and aerial systems which may be used for telephony are described in the appropriate chapters of this book and we will now deal with the special terminal equipment that is necessary, with privacy methods and with some of the special circuits that are in use.

The Singing Suppressor

The action of a singing suppressor in a 4-wire line is shown schematically in Fig. 371, the terminal equipment at each end consisting of two interlocked suppressors, one in the "go" line and one in the "return." In the quiescent condition both "go" lines are blocked and both "return" lines are clear, as indicated in Fig. 371a.

Speech from a distant subscriber, S_A say, to the terminal at his end, operates the interlocked suppressors there, blocking the "return" line and clearing the "go" line. Since the other end is already clear, the "go" currents will reach the subscriber at S_B, as shown in Fig. 371b. The action is reversed when S_B speaks, but in cases when both subscribers speak together, all four suppressors become blocked, as shown in Fig. 371d. The circuit changes must, of course, be made

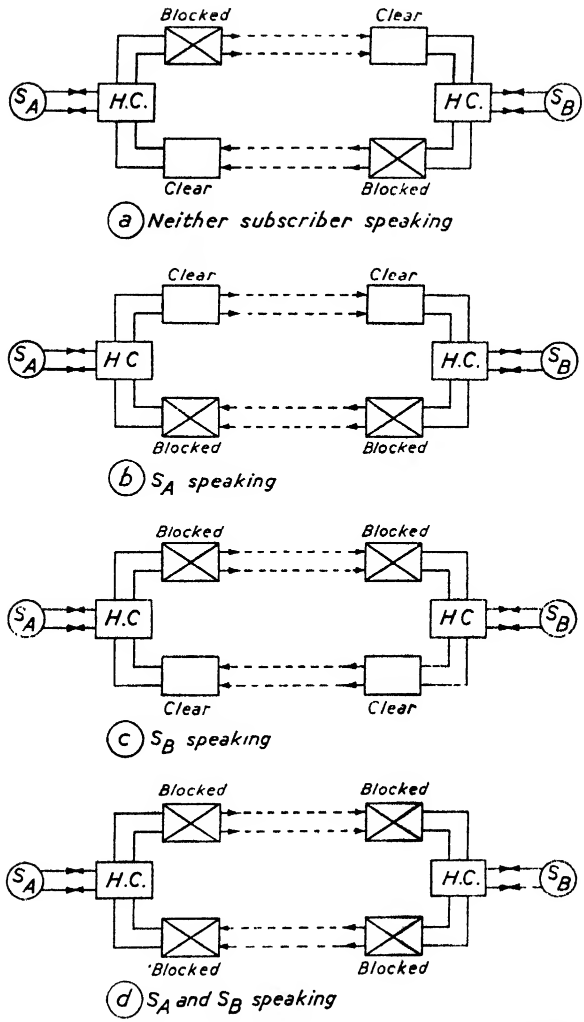


FIG. 371. Illustrating the Action of a Singing Suppressor.

automatically, without the subscribers being aware of anything unusual in the behaviour of the circuit.

The wireless telephone is a special case of such a 4-wire circuit and Fig. 372 shows schematically such a circuit. It will be observed that in addition to possible " singing " around the

main loop circuit, there are also loop circuits at each end from T_1 to R_1 and T_2 to R_2, and should the circuit operate at the same wireless frequency in both directions (which is not, however, usual) " singing " round the local circuit must also be prevented.

The earlier types of singing suppressor employed electro-mechanical relays to switch the circuits and these are still used in the U.S.A. Circuits using valves and others using biassed rectifiers have also been developed. It will be appreciated that the first syllable of speech, usually a consonant, has to switch

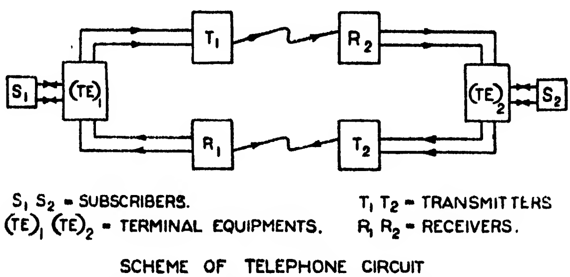


FIG. 372. Wireless Telephone Circuit.

the circuit, but we have seen (Chapter II) that the energy level of most consonants is extremely small. This makes the design of the suppressor difficult and the adjustments critical when the noise level in the circuit is high.

Since the various manufacturers' apparatus has to work over circuits of an international character and conform to a specified performance, there is not a great deal of difference between them and we will describe two types.

British Post Office Terminal Equipment

A simplified circuit diagram is given in Fig. 373. This diagram includes the privacy equipment which requires additional suppressors and an extra hybrid coil, but in the preliminary explanation this will be ignored and we can assume that the input from the wireless receiver is connected to the points *AB* and the speech output to the modulator of the transmitter from the points *CD* direct.

Examination of this diagram shows that it includes two

suppressing networks, RSR_1 and RSR_2, in the receiving line and two suppressing networks, TSR_1 and TSR_2, in the transmitting line. In considering the working of this system we must keep in mind the general conditions set out previously, namely that with no speech passing either way, the incoming line from the receiver is open, and the outgoing line is blocked. Further, incoming speech maintains these conditions but outgoing speech must be capable of reversing the conditions.

The suppressing networks are the means of accomplishing this and each consists of a network of rectifier elements (copper oxide or selenium) to form an attenuator in which the loss can

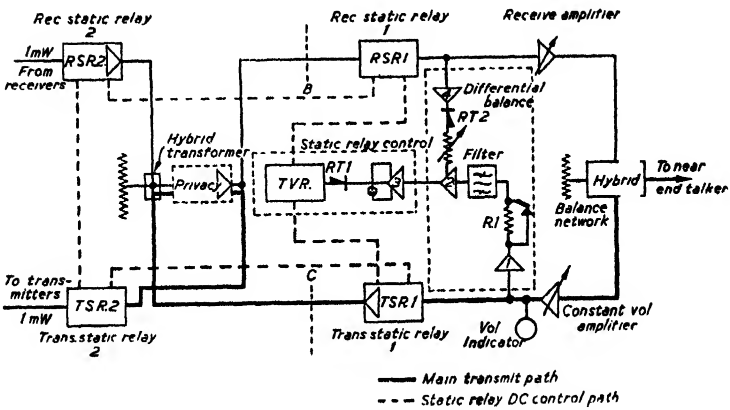


FIG. 373. P.O. Terminal Equipment.

be varied. When the control current, which passes through the four networks in series, is reversed, the loss in RSR_1 and RSR_2 changes from 2 db to 110 db and from 110 db to 2 db in TSR_1 and TSR_2.

The control current passing through the suppressors is produced by the valve relay TVR—a quick-acting device which reverses the current in its output when a voltage, negative with respect to the cathode, is applied to the grid of its input valve. This latter voltage is obtained from rectifier RT_1 and amplifier 3, which in turn is fed from the differential balance.

The differential balance has a special function to perform. It receives signals from both transmitting and receiving lines

and determines whether or not signals present in the transmitting line shall be passed to amplifier 3 and allowed to open the transmitting line.

When no speech is being received from either direction, the output from the differential balance will be zero and the current at the output of the valve relay will be such that the suppressor networks on the receiving side have only 2 db loss and those on the transmitting side have 110 db loss.

When speech is received from the near end subscriber, the signals pass through amplifier 1, filter, amplifier 2, amplifier 3 and are rectified by RT_1. The valve relay operates and reverses the control current, thus closing the receiving line and opening the transmitting line. The bandpass filter (600 to 2,600 c/s) in the differential balance provides discrimination against line noise and the 6 db attenuator R_1 permits a reduction in sensitivity to meet difficult cases.

When speech is received from the distant subscriber (on the receiving line) the signals are amplified by 4 and rectified by RT_2; the voltage produced by RT_2 suppresses 2 which in turn prevents signals from the transmitting line from being amplified and applied to the static relay control circuit.

Received signals now reach the transmitting line and thence the input to 2 by two means, i.e. across the terminal hybrid transformer owing to mis-match between the balance network and the 2-wire line, or by reflections or echoes from impedance irregularities along the 2-wire line. The suppression of 2 must therefore be maintained for a sufficient time after the signals on the receiving line have ceased, for the echoes to have died away. The time for which suppression is maintained may be increased beyond the normal value when 2-wire circuits having very long echo-delays are in use. This extension of the time during which the control is exercised is determined by the voltage across a condenser which is arranged to be charged quickly but to discharge slowly. The quick charge is necessary to prevent signals reaching the input to 2 via the hybrid transformer before the control is exercised.

Under good radio conditions, noise on the receiving line does not produce any voltage at the output of RT_2. Such conditions apply for a considerable part of the scheduled hours of operation. However, at times relatively high noise levels have to

be accepted at the receiving line. In order that the gain of 2 shall not be affected by these high values of radio noise during the periods when there is no speech on the receiving line, provision is made to reduce the sensitivity of the receiving side of the differential balance by back-biassing rectifier RT_2, when the radio noise exceeds a predetermined value.

This expedient, however, results in a smaller suppressing bias applied to 2 when speech signals are received and there is, therefore, a greater danger of false operation of the device by the signals which have travelled from the receiving line to the transmitting line by either or both of the methods mentioned above. The liability to false operation in this manner depends upon the gain in the amplifier in the transmitting line (constant volume amplifier). If the near end subscriber's speech level has resulted in a low gain of this amplifier and the level of the echoes is not exceptionally high, false operation will not occur, but if the speech level has resulted in a high gain and the level of the echoes is also high, some false operation will occur. The difficulty arises with any radio-terminal equipment but the differential balance has the advantage of reducing to a minimum the number of cases in which the further extreme measure of decreasing the gain in the main receiving line has to be adopted.

Singing Suppressor—Marconi Type [6]

This equipment is in use at the various terminals in the British Empire with which the British P.O. conduct telephone services, and, therefore, has to work in conjunction with the P.O. suppressor just described and in principle it is very similar, but a brief description of certain parts may be of interest. A schematic diagram is given in Fig. 374. If privacy equipment is in use an additional hybrid coil and suppressors will be required as in the P.O. type.

Valves having independently heated cathodes are employed although D.C. filament heating is used, because convenient methods of applying bias voltages are thereby rendered possible.

The differential principle is used, but the means of applying it are quite different, as will be seen from Fig. 375. One of the inputs has come from the receiving side and the other from the

transmitting, each having come through an amplifier gain control and filter. Each input undergoes full-wave rectification so that uni-directional currents flow in the resistances *OP*, *OQ*, both currents flowing away from the centre point. If the

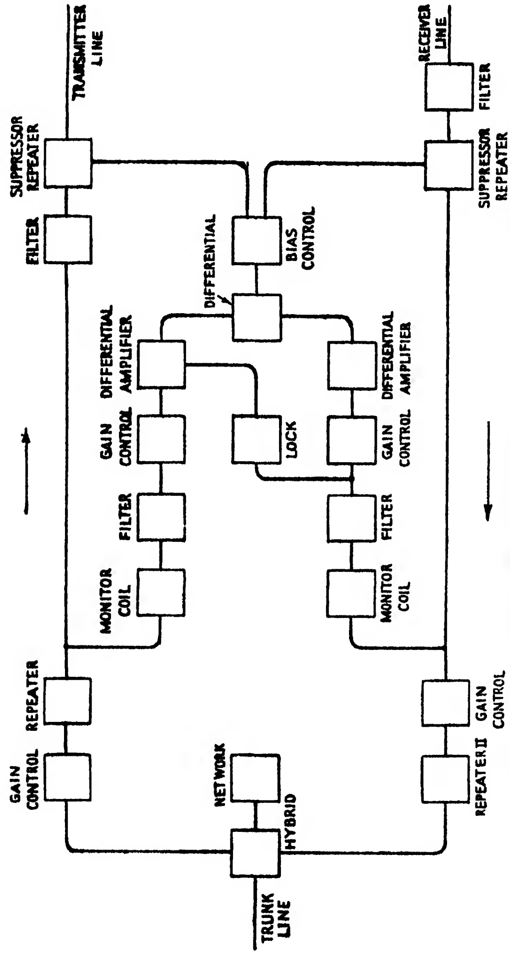


Fig. 374. Block Diagram of Marconi Terminal Equipment.

inputs are equal (and the valves matched) there will be no voltage across *PQ* but any inequality of input produces a D.C. voltage which is applied at the grid of the first valves in the bias control circuit. If there is no output from the differential, V_1 is cut off and there is no P.D. across R_1 to bias the receiving

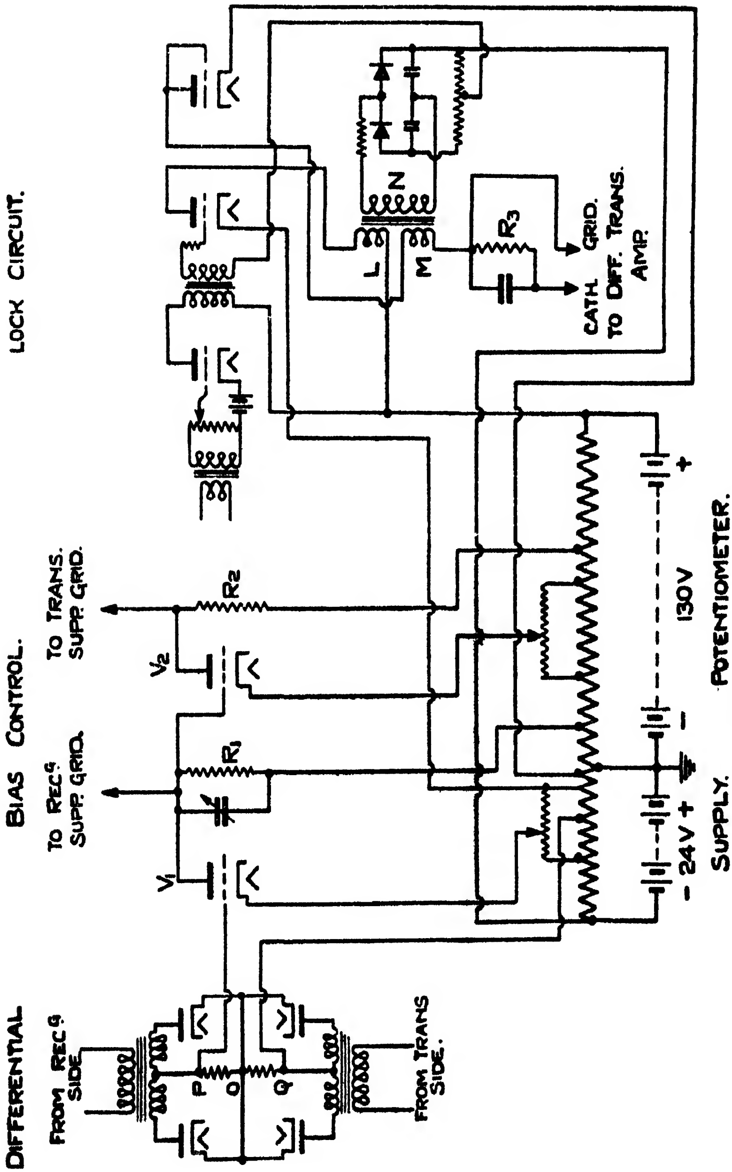


FIG. 375. Details of Marconi Terminal Equipment.

suppressor, but the grid of V_2 is positive and hence there is a P.D. across R_2 to bias the transmitting suppressor. If the receiving input becomes greater than the transmitting input, V_1 is merely made more negative, but if the transmitting input becomes the greater V_1 becomes conducting and if the disparity is sufficient the receiving suppressor will be blocked and the transmitting suppressor opened. Owing to the condenser the change back to the receiving condition takes a time to be effected, which may be varied.

The above differential arrangement has an advantage over circuits in which a grid bias is balanced against a speech input. Because the inputs from the transmitting and receiving sides are of the same nature, the differential is linear, that is, a given difference between the two inputs always produces the same P.D. across the output. In consequence of this, re-adjustment is not necessary if the noise level changes by a large amount.

A special feature of the equipment is the " Lock " Circuit, also shown in Fig. 375. This is so arranged that when the input from the receiving side exceeds a certain value the transmitting circuit is locked in the suppressed condition, irrespective of what voltages may be impressed on the differential from the transmitting side. The " Lock " is, therefore, so adjusted so that it does not work on received noise but does work on received noise plus received speech.

It will be seen that the first valve of the " Lock " Circuit is a straightforward amplifier taking its input from the receiving side whilst the second valve has a three-winding transformer in its anode circuit. Considering only the windings L and M it will be seen that as the input increases, the voltage applied to the diode increases, but because it is negatively biassed, the current in R_3 rises quickly after a certain input is reached. This effect is enhanced by the arrangement of metal rectifiers connected to the N winding which removes the bias on the second valve as the input increases, slowly at first, and then rapidly for larger inputs. The net effect is that at a certain value of input a large P.D. is produced across R_3 and this paralyses the differential transmitting amplifier. The condenser across R_3 ensures that the transmitting side remains suppressed until any echoes of the received speech returning from the 2-wire circuit have been dissipated.

Some special methods of working will next be discussed before considering the question of privacy because some of these special methods give, incidentally, a measure of privacy.

Suppressed-Carrier Single Sideband System

The theory of this method of transmission, referred to simply as S.S.B., has been fully dealt with in Chapter III, and the limitations discussed. The method has been employed on the long-wave transatlantic telephone circuit since its inception in order to economise in transmitter power and reduce the frequency band occupied. Although these considerations do not press so heavily on short-wave telephone circuits and the difficulties are much greater due to the higher frequencies, a suppressed carrier system is less liable to distortion when selective fading is present (see page 153) and the advantages to be gained by its adoption are so great that international organisations are changing over to S.S.B. completely.

It will be evident that if the carrier fades deeply whilst the sidebands remain at full strength the effect on the detector is the same as over-modulation and large second harmonics will be produced in the detector output. Any ordinary system of gain control depends upon the carrier and is, therefore, likely to increase the distortion when selective fading of the carrier occurs by causing the sidebands to be amplified unduly. Considerable improvement in quality might, therefore, be expected if the varying carrier were replaced by a steady carrier generated at the receiver. It is especially important, however, to prevent the introduction of spurious frequencies when privacy systems are in use.

Let us compare the signal/noise ratio obtained from an ordinary telephone transmission and from a suppressed carrier single sideband transmission. The basis of comparison and the assumptions made must be clearly understood or the figure arrived at is meaningless.

As a basis we will take it that the peak voltage in the final stage of the transmitter is the same for the two cases. When water-cooled valves are used, the limit of output will normally be the peak emission which the valve can give (see Chapter X), and hence by comparing on the basis of equal peak voltages

we are really assuming that the final stage of the two transmitters will be the same.

The following assumptions are made :

(*a*) Sine wave modulation, the ordinary transmitter being 100% modulated.

(*b*) A distortionless transmitting path.

As an example, let the maximum value of the carrier voltage in the ordinary transmission be 10,000 volts, then the peak value with 100% modulation will be 20,000 volts and the maximum voltage of each sideband is 5,000. In the single sideband transmitter, however, the maximum value of the sideband will be 20,000 volts or four times that of each sideband in the ordinary transmission. If the transmission path is distortionless then the two sidebands of the ordinary transmission will arrive in the correct phase relationship and add their effects at the receiver detector. Hence the output voltage from the single sideband receiver will be twice that from the ordinary receiver, the power output being four times, so that the gain is 10 log 4 or 6 db approximately. The single sideband receiver need have only half the bandwidth of the ordinary receiver and hence the noise energy picked up will be approximately halved. The signal/noise ratio in the S.S.B. receiver will therefore be 10 log 8, or 9 db above that in the ordinary receiver, 6 db of this being due to the more favourable transmitter conditions and 3 db because of the halved bandwidth. If we take into account the calculated gain of 3 db due to less selective fading and no phase displacement between the sidebands, we get a total gain of 12 db. In practice the changing over of an existing channel to S.S.B. working would mean that the transmitter power could be greatly reduced during almost the whole working time.

The S.S.B. system has the disadvantage that it does not in itself provide a sufficient measure of secrecy and the simpler form of privacy (speech inversion with carrier wobble), which can be used with the double sideband system, is not effective.

We showed in Chapter III that for a S.S.B. system to be effective the receiver must either know the exact frequency of the suppressed carrier or must be able to judge when the reintroduced carrier is correct by the type of signal received.

If simple speech is sent over a channel, then the second criterion is effective. If privacy systems are in use, or if the circuits are being used for voice-frequency telegraphy, then this criterion fails and an exact knowledge of the suppressed-carrier frequency would be needed.

It is clear that the permissible variation in frequency between the suppressed carrier and the reintroduced carrier will be the same number of cycles whether the carrier frequency is low or high, and hence the special difficulty in a short-wave application of the method will be the very small percentage variation permissible in the reintroduced carrier. For good quality it is considered that 20 c/s for speech and 4 c/s for music represents the maximum tolerance permissible.

Such close agreement between independent R.F. oscillators would be exceedingly difficult, if not impossible, to maintain and no such system has been tried out. Since independent carriers at transmitter and receiver are not practicable, it becomes necessary to transmit some pilot signal or carrier which can be used to control the receiver oscillator, so that small variations in the transmitter frequency are automatically passed on to the receiver.

A number of alternatives present themselves, three of which have been the subject of extensive tests.

(*a*) The carrier and one sideband may be as completely suppressed as possible and a " pilot signal " applied at a low level as part of the modulation at the transmitter, this pilot signal being used to synchronise a local oscillator which reintroduces the carrier.

(*b*) One sideband may be suppressed and the carrier transmitted at a very low level relative to the sideband. After separate treatment at the receiver, the carrier may be reintroduced at the correct level.

(*c*) Transmission as in (*b*), but the carrier is used at the receiver to synchronise a local oscillator which reintroduces the carrier.

Methods (*b*) and (*c*) are now usual, and with most S.S.B. circuits provision is made so that two separate intelligence channels can be transmitted. Although each, of course, will be a S.S.B., neither need be telephony, in fact one channel anyway may often be used for some form of telegraph code

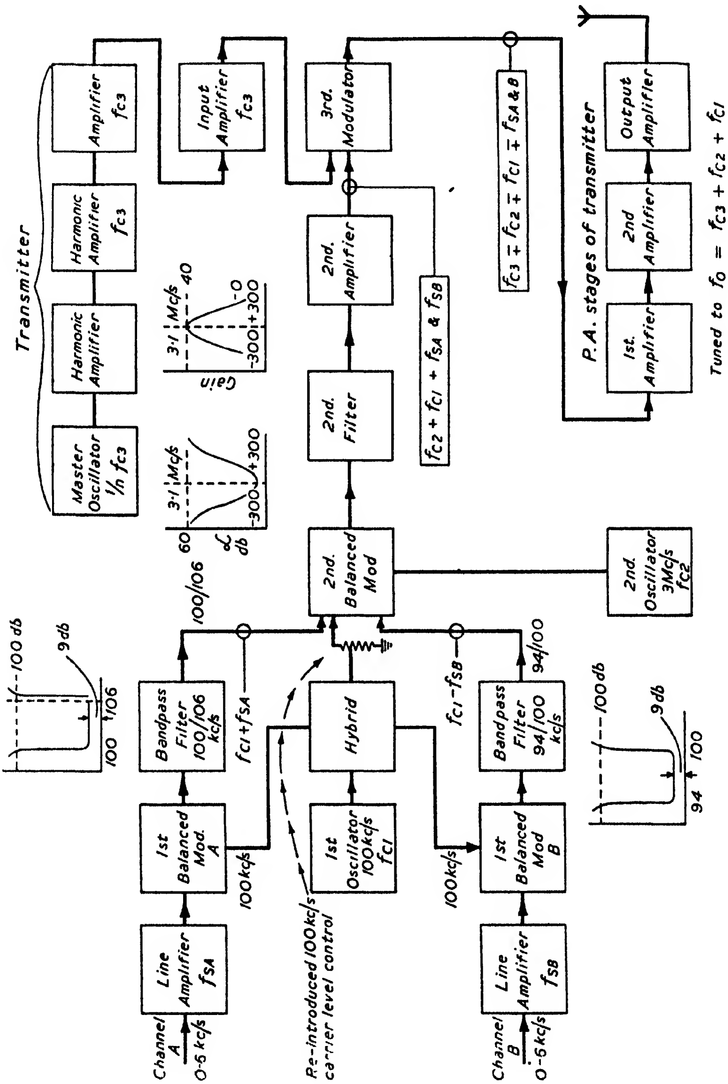


FIG. 376. Dual-channel S.S.B. Transmitter.

operated with voice frequency apparatus as will be explained in the next chapter.

Dual-Channel Single Sideband System

The following descriptions apply to the circuits used by the British Post Office, but as the system is international in character it is typical of those in use generally.

Fig. 376 shows a simplified block schematic diagram of the transmission equipment for a dual-channel circuit from which it is seen that there are three frequency changes made before the final radiation frequency f_o is obtained. The first two changes, at the first and second balanced modulators, are fixed as shown at 100 kc/s and 3 Mc/s, but the third is determined by the frequency to be radiated. This means that to change the frequency of transmission it is only necessary to alter this third oscillator frequency, f_{c3}, in addition, of course, to altering the tuning of the circuits of the transmitter proper.

Referring to Fig. 376 and considering only Channel A, intelligence from this channel is passed to the differential winding of the first balanced modulator, into whose parallel input is fed the first oscillator frequency of 100 kc/s.

If perfectly balanced, the differential output will contain both sidebands and no carrier but in practice a small residual carrier is usually present. These frequencies are passed into the bandpass filter having the characteristic shown in Fig. 376, which passes the upper sideband and suppresses the lower, and this upper sideband is then passed to the differential input of the second balanced modulator together with the original 100 kc/s carrier, which is thus reintroduced here at a controlled level as indicated, much below the sideband level. The output from the second modulator, together with the residual 100 kc/s carrier, is then passed into the filter-amplifier indicated, whose characteristics are such that the upper sideband is again accepted with the residual carrier, and these are then passed to the input of the third modulator, into which is fed the third oscillator. This modulator is not balanced as the resulting sidebands are so far apart from each other and from the third carrier that the transmitter tuned circuits can easily filter out the necessary

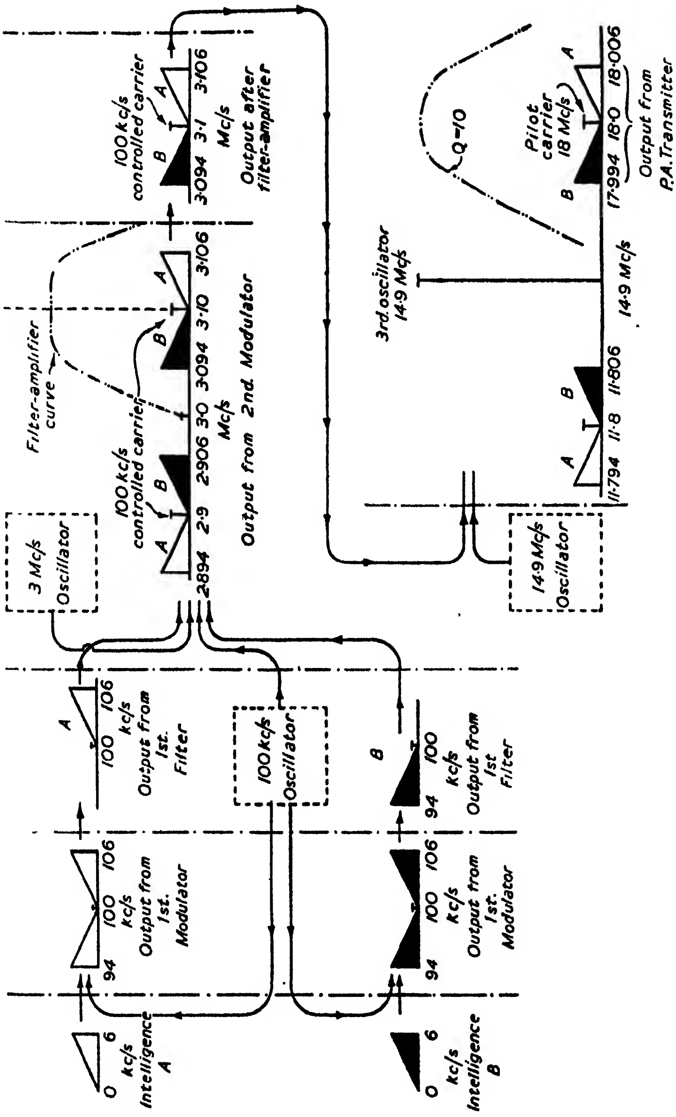


FIG. 377. Illustrating Operation of S.S.B. Transmitter.

single sideband and also f_{c3}. As explained, the third oscillator frequency is determined by the final required frequency.

It will be rather easier if we take actual figures for the frequencies involved so as to see how the correct S.S.B. is selected at each stage and unwanted frequencies eliminated. We will assume that channel *A* intelligence requires a bandwidth of 6 kc/s from zero—the maximum that is possible with the equipment. Since the first oscillator is 100 kc/s, sidebands of 94 to 100 and 100 to 106 kc/s are produced, of which only 100 to 106 pass to the second modulator input, together with the reintroduced 100 kc/s carrier. Emerging from the second modulator, whose carrier is 3 Mc/s, we have sidebands of 3·100–3·106 and 2·894–2·900 because of the signal; and sidebands of 3·100 and 2·900 because of the 100 kc/s reintroduced carrier. One or two other very low level products also come through, but they can be neglected. Because of the characteristics of the filter and amplifier following the second modulator, which are indicated on Fig. 377, only the reintroduced carrier at 3·100 and the upper sideband of 3·100–3·106 Mc/s are passed to the next modulator stage.

We have now arrived at the point where the final frequency change is made, and the value of the third oscillator will be determined by the required radiated pilot carrier and frequency band. If the frequency of the pilot carrier is f_o then f_{c3} will be equal to $f_o - (f_{c2} + f_{c1})$. Thus if we wished to use a final carrier of 18 Mc/s, then the third oscillator could be either 14·9 or 21·1 Mc/s. Assume we choose 14·9 Mc/s, this is passed to the third modulator with the incoming frequency band and an unbalanced modulator is now used as the 14·9 Mc/s oscillator frequency, and the sidebands are so separated that simple tuned circuits can select the required band without difficulty. This increasing separation of the sidebands is shown in Fig. 377, which indicates the approximate filter characteristics at each stage and the elimination of the unwanted frequencies.

Fig. 377 shows both channels, but the intelligence from channel *A* is shown plain.

That part of the circuit up to the output of the second I.F. amplifier is S.S.B. input equipment but the circuits beyond belong to the transmitter proper, and are usually obtained by small modifications to normal wireless telephone transmitters,

Thus, considering a *normal transmitter* to consist of a M.O. with harmonic amplifier and a series of power amplifiers, the M.O. and harmonic amplifier will be detached and utilised as the third oscillator input f_{c3}, shown in Fig. 376, fed into the modulator through an appropriate network. The remaining portion of the transmitter, i.e. the power amplifier stages proper, are supplied from the output of the third modulator as shown in the figure, but it is essential for all these stages to be strictly linear in action, and if not already designed for linear working they must be modified by any of the methods described on page 389.

Let us now consider the second part of the dual system, which, as can be seen from Fig. 376, links up with a common circuit at the input of the second modulator. The only difference between the first channel *B* section to that of channel *A* up to the common point, is that the band-pass filter of *B* passes frequencies from 94–100 kc/s, whereas *A* passes 100–106 kc/s. Thus, let us consider incoming intelligence to channel *B* to comprise the same frequency band as coming in at *A*, namely from 0 to 6 kc/s, but to avoid confusion in the sideband diagram of Fig. 377 we will show this intelligence band blacked in, as distinct from the plain band from *A*. The first balanced modulator contains both sidebands as does that of the *A* channel, but whereas the *A* filter passes the upper sideband to the second modulator, the *B* channel passes the lower. Hence, if both *A* and *B* channels are on simultaneously, we have, in addition to the controlled 100 kc/s carrier, the upper sideband from *A* and the lower sideband from *B* passing into the second modulator. Each of these produce, with the 3 Mc/s oscillator, two sidebands, but as can be seen from Fig. 377 only one of each will pass through the second filter amplifier stages, and thus we still feed into the third modulator one sideband of each together with the controlled carrier. Here again, although two sidebands are produced from each intelligence, the filtering action is sufficient to separate one pair, leaving us with an output of a pilot carrier together with the upper sideband of *A* and the lower sideband of *B*, or vice versa, depending on the value of third oscillator frequency chosen, these various separations being shown in Fig. 377.

The power output of a double sideband transmitter is specified by giving the power radiated in the carrier but this is clearly impossible with a single sideband transmitter. The agreed method of specifying the power is by stating the peak power, that is, the output when the full modulation which is permissible without overloading has been applied. It follows that the rating of the 60 kW amplifier already mentioned would be about 15 kW if it was employed as a double sideband transmitter.

Two main factors determine the level of the pilot carrier. Too high a level increases the noise ratio, the degree of privacy is reduced, cross-talk between channels is increased, and the power gain of the circuit is reduced. If, however, the pilot has too low a level, it may fall below the receiver noise level at times and be useless either for heterodyning or frequency-controlling. For normal conditions of working the level of pilot to peak sideband power is about −16db when a single channel is in use and −26db when both are operative. Of course any discontinuity of carrier due to very adverse propagation conditions may necessitate temporary manual retuning at the receiver because of the narrow bandwidth of the carrier reconditioning circuits that are used at the receiver.

It is observed that the circuits have a 6 kc/s bandwidth, i.e. twice the width necessary for a telephone signal, so that theoretically two telephone channels each side, one pair having a displaced frequency band, or an equivalent number of voice operated telegraph circuits, could be handled. But to do this the apparatus both at the transmitter and receiver will have to have very linear characteristics, since any non-linearity will introduce new frequencies and in consequence cross-talk, actually of an unintelligible kind. On account of this difficulty and because it is worse the nearer the intelligence bands are together, dual channel working was often used with one intelligence band displaced from 0 to 3 kc/s to 3 to 6 kc/s, so as to reduce the cross-talk. The frequency band 0 to 3 kc/s may then be used for a low-grade order wire circuit, that is a circuit used by the control operator. Of course, this increase of bandwidth reduces the signal/noise gains that have been assumed. It has now been found possible to work without this separation although the power put into each channel at the transmitter

has to be reduced about 6 db below that used when one channel is in use. It would in any case be necessary to reduce the power in each channel by 3 db if the transmitter peak power is to remain the same.

A block schematic of the Post Office Dual Channel Single Sideband Receiver is shown in Fig. 378. It is seen to be a double superheterodyne, having I.F.'s of 3·1 Mc/s and 100 kc/s so that some of the filters are identical in design with those in the transmitter.

The first oscillator is usually a crystal oscillator when the receiver is being employed on fixed services.

After amplification at 100 kc/s, the A and B channels and the

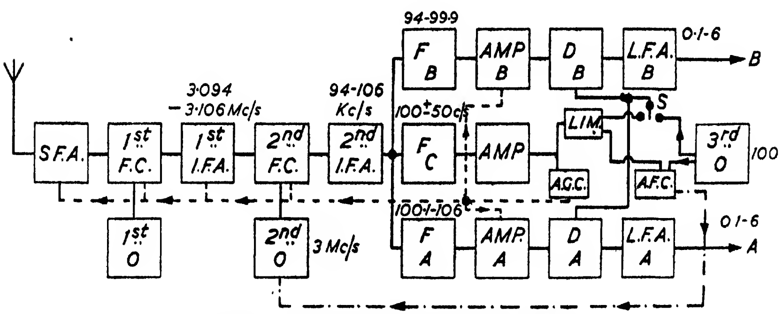


FIG. 378. Block Schematic of S.S.B. Receiver.

carrier are separated out by filters F_A, F_B and F_C. The bandwidth of the carrier filter, F_C, at 6 db down, is 100 c/s and the attenuation at 250 c/s from the carrier frequency is 60 db. It is desirable that the pass band of F_C should be as narrow as possible in order that the signal/noise ratio of the low-level carrier may be good enough, but if it is made too narrow the A.F.C. might fail if there was a sudden change in the frequency of one of the oscillators in transmitter or receiver, even if the change was very small.

After the carrier has been amplified, it is used to operate the A.G.C. It will be seen that the bias derived from the A.G.C. is applied to most of the valves *preceding* the point at which the carrier is separated, in the usual way. The characteristics of the A.G.C. are such that an increase of 80 db in the input signal to the receiver would produce about 6 db increase in

the channel outputs. This is reduced by applying a portion of the A.G.C. bias to the channel amplifiers *following* the point at which the carrier is separated.

The carrier can be reintroduced by either of the methods (*b*) and (*c*) given on page 623, depending upon the position of *S*. In either case the carrier is passed through a limiter and then to the A.F.C. Two types of A.F.C. have been employed, one being an elaboration of the discriminator and reactance valve circuit often used in less complex receivers. In such a system a certain change of frequency must occur before a sufficient voltage is produced to effect the correction, that is, there must be a residual frequency error. By incorporating a crystal resonator in the discriminator circuit, a frequency drift of 4 kc/s can be corrected with a residual error of 10 c/s.

The alternative, and probably superior, method is electro-mechanical. The output from a 100 kc/s oscillator is mixed with the nominally 100 kc/s carrier in a system of modulators and phase shifters, so that a 4-phase supply at the beat frequency is obtained. That is, currents at the beat frequency are produced which have relative phases of 0°, 90°, 180° and 270°. These are passed through four windings spaced at 90° round the stator of a small motor and consequently produce a rotating magnetic field which makes one revolution for one cycle of the beat frequency. The rotor therefore revolves at the beat frequency and this drives, through gearing, a variable condenser in the resonant circuit of the second oscillator. The beat frequency is therefore reduced until, when it becomes zero, the motor stops. Should the frequency drift in the opposite sense, the rotating field will revolve in the opposite direction and hence the frequency will again be corrected.

It will be seen that this type of frequency correction need have no residual frequency error, since the smallest difference in frequency may start the motor revolving very slowly.

The "Ray-Diversity" Receiver

In Chapter IX we mentioned the Multiple-Unit Steerable Array and it is now necessary to consider the type of receiver to be used with it. As the receivers used are somewhat complex, space only permits of us giving the barest outline and the paper by Gill [12] should be consulted for further informa-

tion. Referring to Fig. 379 which shows a greatly simplified schematic diagram it will be seen that the sixteen rhombic arrays each have their own R.F. and 1st I.F. circuit. At the second frequency-changer, arrangements are made to correct the phase of the inputs from the 16 aerials so that they may be combined. This is done by supplying local oscillator voltages differing in phase. The phase-differences are obtained by an

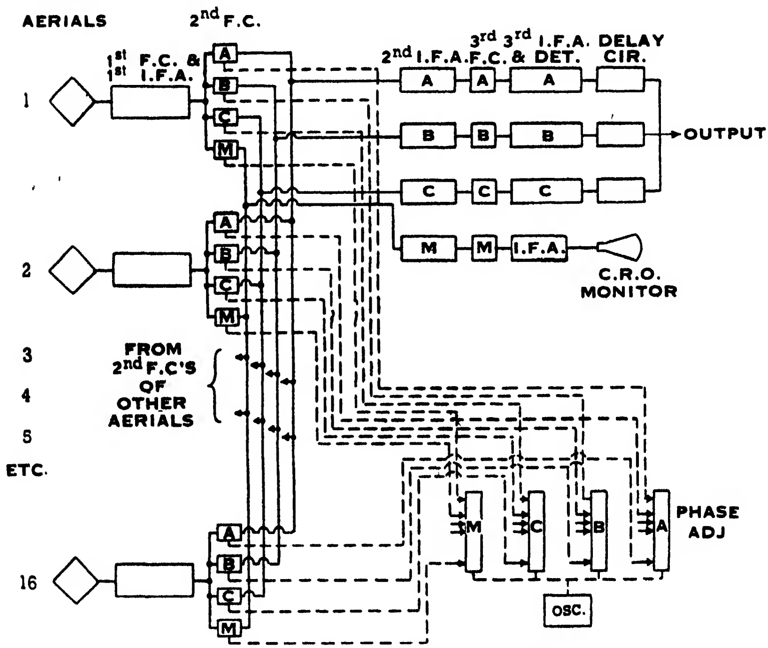


FIG. 379. Ray-diversity Receiver.

ingenious arrangement of artificial line and auxiliary oscillator (detail not shown in Fig. 379), the adjustment being the frequency of the latter.

Four separate phase-adjusting circuits are available so that four separate zenithal polar-diagrams can be obtained at the same time. Hence from the second frequency-changer there are four separate circuits *A*, *B*, *C*, and *M*, each corresponding to a certain (adjustable) reception angle. Three of these circuits, after further amplification and rectification, are combined at audio frequency as indicated.

Thus if the energy is reaching the receiving aerials by a number of rays, the three best of these can be selected and the energy finally added. Delay networks can be inserted at audio frequency to allow for the fact that the ray paths may be of different lengths. If there are less than three rays the unwanted circuits may be cut off.

The fourth path M through the receiver is the monitoring circuit which terminates at a cathode-ray oscillograph. The oscillator phase of this path is being continually varied so that its zenithal angle of maximum reception swings over a range of values periodically, and an inspection of the oscillograph screen enables one to see at a glance at what angles useful rays are coming in and to adjust the A, B, and C paths accordingly.

The actual receiver is for dual-channel, single sideband working and a number of receivers could be connected to the same array system.

The M.U.S.A. system has been in operation at both ends of the London–New York circuits since 1942 and has improved the service. It has been found more satisfactory to find out, with the aid of the monitoring channel, which ray is giving the strongest and steadiest signals and then to adjust one of the channels to receive this—the other two channels would not then be used.

Quiescent Carrier Systems

It is evident that a considerable saving of power can be effected if a telephone transmitter (which is radiating the carrier and both sidebands) has its carrier emission stopped whenever speech is not actually being transmitted. Measurements made by the Post Office indicate that during an ordinary conversation each transmitter is only transmitting speech for about 13% of the time and there will be many pauses between conversations when neither transmitter is actually in use.

In addition to the economy effected, " eavesdropping " becomes more difficult, since the carrier is only transmitted for brief periods.

When one end of a telephone circuit is on shipboard it is almost essential for the transmitter on the ship to work on the quiescent carrier principle because of the very small separation

between transmitter and receiver. Comparatively heavy currents are induced in the metallic structures of the ship. These structures are rarely "linear" conductors and thus give rise to a vast quantity of intermodulation products, which cause noise in the receiver.

It is expected that the telephony service to the large liners will, before very long, be worked on the single-sideband system.

One method of achieving carrier suppression is by a modification of the absorber keying circuit described on page 535. The incoming speech is fed into the input of a control amplifier, Fig. 380, as well as into the line amplifier which modulates the transmitter. The control amplifier is followed by an anode

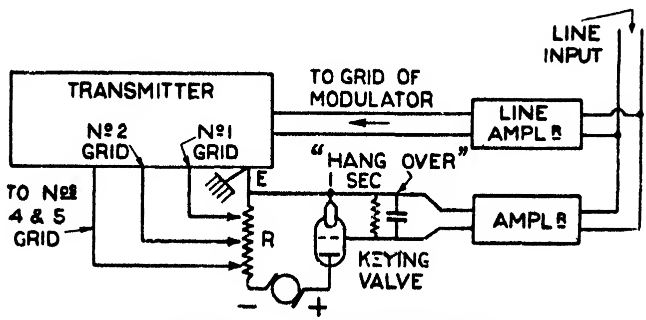


FIG. 380. Quiescent-carrier Circuit.

bend rectifier (included in the amplifier in Fig. 380), the D.C. output voltage of which is applied to the grid of a valve which takes the place of the keying relay.

With no speech, since the rectifier valve of the control amplifier is then passing no feed, the grid of the keying valve is at positive potential and a heavy feed flows through the keying resistance, causing the fourth stage to be cut off, and thus preventing radiation of the carrier. When speech is applied the rectifier takes feed, the keying valve grid is backed off negative, the current through the resistance is stopped, and the circuit drives in the normal manner.

It will be realised that it is not desirable such a circuit should cut off the carrier in the intervals between syllables of speech, and to prevent this a time constant circuit (shown in Fig. 380) is provided whose constants can be arranged such that only for "pauses" will the carrier be suppressed.

A difficulty is introduced at the receiver by the employment of quiescent-carrier working at the transmitter. The normal type of automatic gain control depends upon the carrier for its operation. If, therefore, the carrier fails, the gain control will push up the amplification to the extreme limit, thereby raising the noise at the receiver output to a very high level. Since no speech is being received at the moment, this might not appear to matter, but this noise passing along to the terminal equipment is liable to lock the suppressors in the "receive" connections and prevent transmission.

One solution is the use of a similar arrangement to that employed in some broadcast receivers in order to prevent an irritating noise from the loudspeaker when tuning between stations. All the stages of the receiver following the gain control are kept paralysed unless a signal of a certain minimum strength is picked up. This is more difficult to do on a commercial short-wave receiver because the minimum signal strength it is required to receive is so much smaller than in the case of a broadcast receiver.

The Compandor System [9]

The name of this system is a compound of compressor and expander because at the transmitting end of the circuit the amplitude range of the signal is compressed, whilst at the receiver it is expanded again. The object is to improve the signal/noise ratio and, as this is poorest when long waves are used, the method is only used at present on the long wave transatlantic circuit but is included here for completeness.

In order to obtain the best signal/noise ratio it is evidently necessary that the transmitter should always be fully modulated. The strength of speech arriving at the terminal equipment from the 2-wire circuit naturally varies greatly with the type of connection and peculiarities of subscriber. It is estimated that on an ordinary telephone system there is as much as 70 db difference of level between a consonant as rendered by a soft-voiced speaker and a vowel spoken by a loud-voiced one.

Manual adjustment at the terminal equipment is always resorted to in order to adjust for widely differing subscribers, and this reduces the amplitude range to something like 30 db

but cannot, of course, allow for sudden changes or the very different energy levels of different syllables. It follows that on a circuit where the signal/noise level is poor, if the highest level portions of speech fully modulate the transmitter the lowest level (but very significant) portions may fall below the noise level.

In the Compandor System the difference in level is reduced to about 15 db by the compressor and then expanded at the receiver, usually to the original amplitude range.

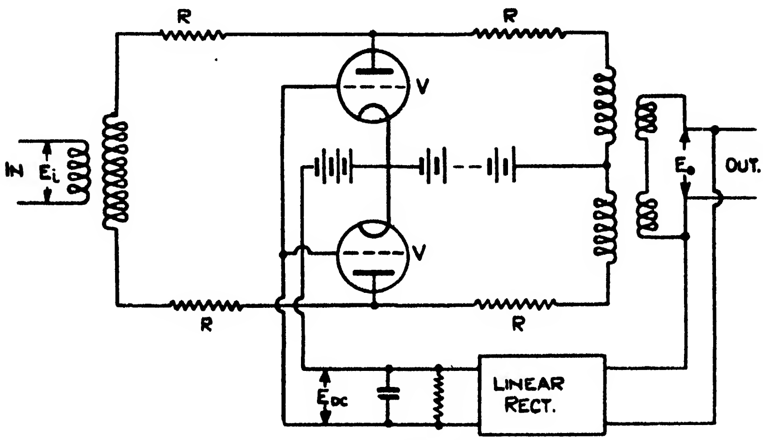


FIG. 381. Compressor Circuit.

The circuit of the compressor is shown in Fig. 381. The valves $V V$ are used as variable resistances and it will be seen that if their resistance is decreased, the voltage available at the output for a given input will be decreased. The valves are set on the curved portions of their characteristics so that R_B, the anode-filament resistance, is inversely proportional to E_{DC}.

If the output voltage at a given time is E_O, then this may be written K_1, E_i R_B where E_i is the input voltage whilst K_1 is a constant dependent upon the transformer and resistances R R R R. E_{DC}, the value of the D.C. output from the linear rectifier, is given by K_2 E_O and

$$R_B = \frac{1}{K_3 E_{DC}} \qquad \frac{1}{K_3 K_2 E_O}.$$

Substituting this value for R_B in the first equation for E_O gives

$$E_O^2 = \frac{K_1 E_i}{K_3 K_2} \text{ or } E_O = K\sqrt{E_i}.$$

Hence, if the level of E_i varies from time to time by 30 db the level of E_O will vary by only 15 db.

The resistance-condenser combination in the rectifier output has a small time-constant so that the compressor works at about syllabic frequency.

The expander circuit used at the receiving end is shown in Fig. 382, and it will be seen that the valves are biassed to the

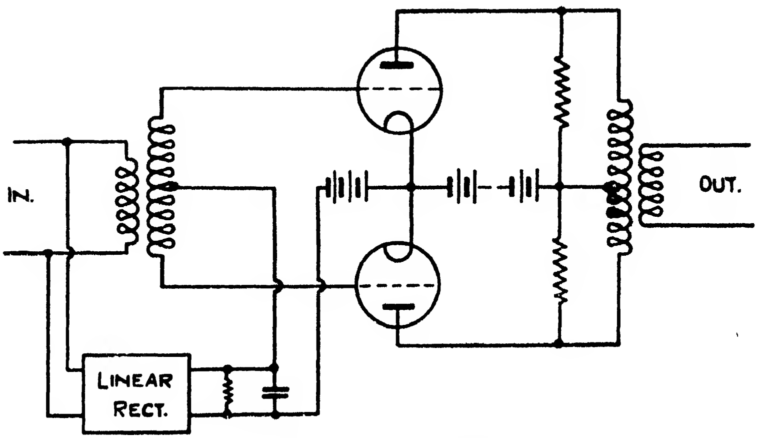


FIG. 382. Expander Circuit.

curved portions of their characteristics and that a portion of the input is rectified and reduces this bias. The amplification of the received signal will, therefore, be approximately proportional to the amplitude of the input and hence the amplitude compression of the transmitted speech will be removed.

It will occur to the reader that the use of the curved portions of the valve characteristics is likely to result in the production of unwanted even harmonics. This is prevented by the use of push-pull circuits.

Privacy Systems—A General Discussion [10]

Although nothing seems easier than the accidental production of unintelligible speech, yet when we wish to render speech

unintelligible to all except one special receiver the matter becomes difficult, and to give successful results requires an elaboration of apparatus.

The special methods of working which have been discussed in the previous section do provide in themselves some measure of privacy if undistorted speech is applied to the transmitter, as any one equipped with simple receiving apparatus would need to have very fine adjustments if they wished to pick up the transmission with any degree of intelligibility. These methods, however, can hardly be regarded as secret systems in themselves.

The privacy systems most used are voice frequency systems situated at the terminals of the circuit and are, in fact, equally applicable to line working if desired.

By suitable circuit arrangements speech may be "inverted," that is low frequency components of the speech become high, and vice versa, the inversion frequency point being usually 1,500 c/s. Simple inversion by itself is not of much value and it is usual to incorporate with inversion a frequency change of carrier which can only be cleared by a properly designed receiver employing inversion equipment. Such a system is suitable for and is used with D.S.B. circuits, but is of no value with a S.S.B. since it is not then possible to add the carrier frequency change owing to the synchronising devices used at the S.S.B. receiver. Alternative methods which are suitable for both D.S.B. and S.S.B. systems involve dividing the speech frequencies into bands (by means of sharp band-pass filters) which may be variously treated and then reassembled for modulating the transmitted carrier. We may, for instance, change the frequency of some of the bands, invert some and not others, and reassemble. Such a system is known as "Split Band Privacy." Or we may introduce varying delay times into the various bands so that the syllables of speech are transmitted in the wrong time order, and are then shifted back at the receiver. Such a system has the disadvantage that the transmission time is increased. Of these methods, although simple inversion with frequency changes is still used with D.S.B. circuits, the "Split Band Privacy" method is becoming increasingly employed and more or less standardised.

Sometimes the subscriber's speech-band is bodily moved

from 0–3 kc/s to 3 kc/s–6 kc/s the band 0–3 kc/s being employed for a low-grade order wire as already mentioned. This displaced band helps in that cross-modulation will be much less in the 3–6 kc/s region due to the fact that most privacy methods employ L.F. modulation stages and the cross-modulation noise will always be worst nearest the carriers.

All privacy methods, even simple inversion, require elaborate apparatus, but the same equipment can be used for speech in both directions, by a suitable arrangement.

Referring to Fig. 373, it will be noticed that additional singing suppressors and a second hybrid are employed. The control for the second receiving suppressor, RSR_2, is linked in parallel with RSR_1 and the control for TSR_2 is linked in parallel with TSR_1. This means that when RSR_1 is clear, so is RSR_2 and TSR_2 will be blocked. This is the condition for incoming speech.

Thus speech from the wireless receiver passes through RSR_2 and from the hybrid through the privacy equipment, which inverts it, and into the terminal in the ordinary way. The connection from the output of the privacy equipment to the suppressor TSR_2 can do nothing, because TSR_2 is blocked.

Considering the condition when transmitted speech is operating the gear, both suppressors RSR_1, RSR_2 are closed and TSR_1, TSR_2 are open. This means transmitted speech passes up through the hybrid to the privacy equipment, is inverted, and the inverted speech passes through the transmitting suppressor TSR_2 to operate the main transmitter.

Privacy by Inversion of the Whole Speech Band

If a speech wave having frequencies extending from 250 to 2,750 cycles be used to modulate a 3,000 cycle carrier, then two sidebands, 250 to 2,750 cycles and 3,250 to 5,750 cycles, are formed. If now we select the lower sideband it will be observed that the speech is "inverted," for it is the 2,750 cycle component of the speech which produces the 250 cycle frequency, whilst the 250 speech component produces the 2,750 cycle frequency. Notice that if we had inverted speech and used this to modulate the same carrier, then selection of the lower sideband as before would lead to straight speech. Thus a

single piece of inverter apparatus could be used for the dual purpose of inverting or re-inverting speech.

Privacy systems utilising this principle have been developed, and are in use, but the simple method of inversion indicated above is not usually adopted at the moment for the higher grade apparatus because of the difficulty of separating the inverted speech from the original speech and from the 3,000 cycle carrier. Instead, a double modulation system is used, for example, the speech may initially modulate a carrier of 13,000 cycles per second, and if the speech frequency is represented as S cycles per second, resultant frequencies of $13 - S$, 13, and $13 + S$ kilocycles per second will be obtained. Of these $13 - S$ and 13 are suppressed, and the remaining sideband, $13 + S$ kilocycles, is used to modulate an oscillator having a frequency exactly three kilocycles per second in excess of the first oscillator, that is, 16 kilocycles.

Thus the output from the second modulated oscillator will comprise frequencies of $3 - S$, 16, and $29 + S$ kilocycles per second ; the two latter are suppressed, and the remaining sideband forms inverted speech. This is again seen to be a low frequency band, and is used to modulate the outgoing high frequency system in the ordinary way.

By such a double modulation system the original speech can be completely cleared from the system, and the design of filters for separating the various sidebands becomes more easily possible.

Such a method of inversion is not entirely secret. If an ordinary receiver of the self-oscillating detector type be adjusted to oscillate 3,000 cycles below the incoming carrier, owing to the selectivity of the receiver circuit, the lower sideband terms will predominate and the interference beats with the receiver carrier will cause inversion, although, of course, a strong 3,000 cycle tone will also be heard.

Interception in this way may be prevented by wobbling the carrier frequency of the wireless transmitter about $\pm$ 500 cycles at a very low rate, usually two or three times a second. This prevents re-inversion by the method just mentioned, but makes no difference to reception by the proper method wherein the carrier is eliminated at the detector as usual, and privacy equipment used to re-invert the speech.

It will be appreciated that the above method of privacy is of no use if applied to a single sideband transmission because merely shifting the reintroduced carrier by (in our example) 3,000 cycles will make the speech normal. Additional privacy by wobbling the carrier frequency is no longer possible unless a synchronous wobble of suppressed and reintroduced carrier can be managed.

Split-Band Privacy

This is best illustrated by an example, shown schematically in Fig. 383. The speech (assumed to have a band-width of

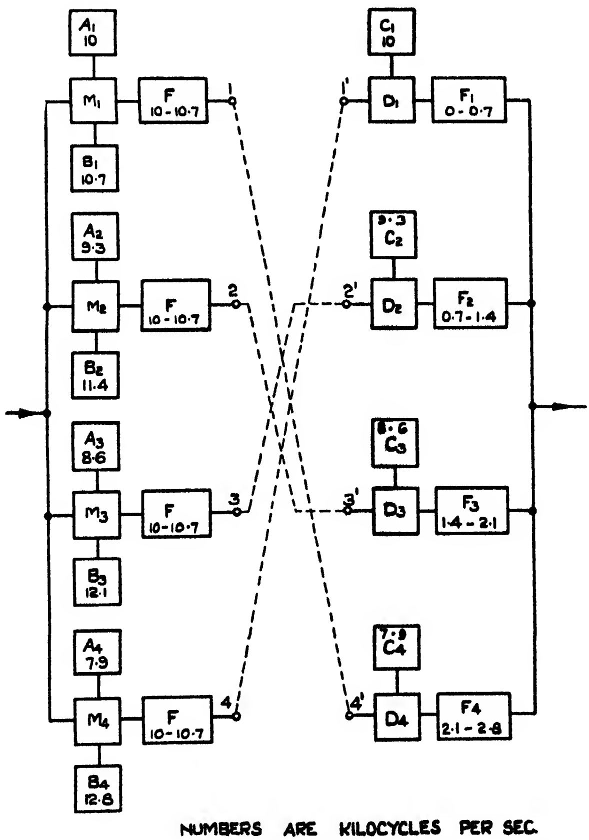


Fig. 383. Split-band Privacy.

0–2,800 cycles) is applied in parallel to four modulators, M_1, M_2, M_3, M_4, each of which combines the speech with the output of *either* its associated *A or B* oscillator. The output of each oscillator is applied to a filter, F, which in every case has a pass band of 10,000–10,700 cycles.

Suppose A_1 is in use, then sidebands 10,000–12,800 and 10,000–7,200 cycles will be produced at M_1, but only the frequencies 10,000–10,700 cycles will pass F. The output from 1 is due, therefore, to those components of the input speech which had frequencies between 0 and 700 cycles and these components are " upright," that is, the lower input frequencies are still the lower frequencies.

Now suppose B_2 is in use. Sidebands of 11,400–8,600 cycles and 11,400–14,200 cycles will be produced at M_2 and 10,000–10,700 cycles only will be passed by F so that the output at 2 is due to components of the speech having frequencies between 700 and 1,400 cycles, but because the lower sideband has been selected these components are " inverted."

In the same way the remaining two circuits will deal with the rest of the speech components, the outputs being " upright " or " inverted " according to whether the A or B oscillator is in use, and all four outputs, each containing different components of the speech, will have frequencies between 10,000 and 10,700 cycles. On this account it is possible to " jumper " any output of this stage of the privacy equipment to any input of the next stage, but in order that an identical equipment at the receiver shall re-convert to plain speech, it is essential that the " jumpering " be complementary to the two equipments, that is, if 1 is connected to 4′ in one equipment, then 4 must be connected to 1′ on the other equipment, and so on. The speech must, therefore, be split into an even number of bands.

The second portion of the equipment consists of four detectors, D_1, D_2, D_3, D_4, each having an associated oscillator, C_1, C_2, C_3 and C_4, the frequencies of which are arranged (as shown in Fig. 383) to produce speech frequency bands between 0 and 2,800 cycles when the inputs to the detectors are all 10,000–10,700 cycles per second.

If the arrangement shown be adopted, the output from 1 will comprise frequencies 2,100–2,800 and a series of distortion terms. The band 2,100–2,800 is selected by F_4 and hence

these frequencies appear in the output due to components of the input between 0 and 700 cycles. Similarly, the other speech frequencies may be followed through the equipment, and it will be seen that the output to the line occupies the frequency band 0–2,800 cycles, but that the relationship between the original and output frequencies is most disordered. An "eavesdropping" receiver requires not only elaborate equipment, but also a knowledge of the "jumpering" and this can be frequently changed.

It is necessary to assure ourselves that an identical equipment will give us clear speech at the output if the distorted speech is fed into the input. Suppose we trace the original speech frequencies between 700–1,400 cycles. These left the transmitting privacy equipment as frequencies from 1,400–2,100 cycles, the band being "inverted." This band of frequencies on reaching the receiving privacy equipment is combined in M_3 and B_3 to form a sideband of 10,000–10,700 cycles which is re-inverted, but still contains the frequency band 1,400–2,100. This is combined at D_2 with the output of C_2 and the frequencies 1,400–2,100 cycles are changed to 700–1,400, freed of distortion by the filter F_2 and pass to line as the original band.

Associated Measuring Apparatus

In order that wireless telephone channels may be efficiently worked it is highly desirable to include in the terminal equipment certain measuring apparatus, so that the performance of the circuit from day to day can be expressed in a more or less

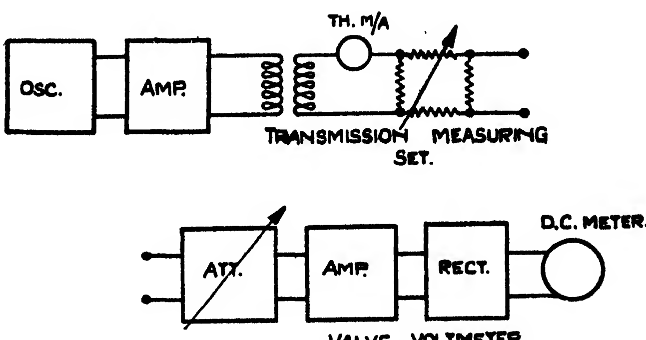


FIG. 384. Transmission Measuring Apparatus.

quantitative way and any preventable deterioration corrected for. The use of such apparatus, indicated schematically in Fig. 384, also assists greatly in the localisation of defects or breakdown.

An oscillator will be necessary as a source of power when taking measurements and will be used for the preliminary "lining up" of the whole circuit.

The oscillator will usually be of the heterodyne type, that is, the output note is produced by combining and rectifying the outputs of two high frequency oscillators, one of these being of large amplitude and the other small. By this means an output nearly free from harmonics is obtained and a wide range of output frequencies may be obtained by a small adjustment of a variable condenser of moderate size in one oscillator.

The oscillator output will be fed through an amplifier having adjustable gain into the transmission measuring set. The first portion of this contains a thermal milliammeter and a network so that this portion can be adjusted to be the equivalent of an alternator having an internal resistance of 600 ohms (and no reactance) and giving an output of 1 milliwatt. Following this is a network of resistances which may be adjusted so that the final output level can be reduced, usually in 1 db steps, the resistance remaining at 600 ohms.

The next part of the measuring apparatus to be considered is the valve voltmeter. The input to this can be arranged to be of high impedance so that it may be bridged across a circuit without disturbing it. A variable, calibrated attenuator is followed firstly by an amplifier, then by a rectifier, the output of which is shown on a moving coil milliammeter.

These components may be used together in a variety of ways. The valve voltmeter is calibrated by producing a known power output from the transmission measuring set and applying this direct to the valve voltmeter. The attenuator is adjusted to give a certain reading on the rectifier milliammeter. If now the loss (or gain) in decibels occurring in a certain portion of the terminal equipment is to be measured, the output of the transmission measuring set will be applied to the valve voltmeter via the apparatus under test and the attenuator adjusted till the rectifier milliammeter reads the same as during calibration. The difference in attenuator settings for test and

calibration evidently gives the loss (or gain) in decibels occurring in the apparatus under test.

If the gain (or loss) in the whole wireless telephone circuit is to be measured then the transmission measuring set at one terminal point will be adjusted to give a known output and this applied at the desired point in the terminal equipment. The received signal will be measured by the valve voltmeter at the other terminal point (after calibration by means of the transmission measuring set at its own terminal).

We have so far only considered measurements conducted when a pure tone is in use but evidently measurements of levels when speech is passing along the circuit will be necessary and it is also desirable we should be able to measure the noise level.

The noise level may be measured on the valve voltmeter in the same way that pure tones are measured, except that the instrument pointer will be rather unsteady due to fluctuations in noise intensity.

In order to monitor speech levels each terminal equipment will include one or more volume indicators. These may consist of a single-stage amplifier having a high input impedance followed by a copper-oxide rectifier and moving coil milliammeter. The milliammeter is scaled in decibels below or above a reference level, the calibration having been obtained with the aid of an oscillator and transmission measuring set, or similar equipment. The volume indicators may be plugged across various points in the incoming and outgoing circuits and, though calibrated by means of a pure tone, give useful direct indications of speech and noise levels. They will not be so accurate as the valve voltmeter previously described since they depend upon the permanence of an amplifier.

The measurement of speech level cannot be made in the same precisely defined units as the pure tone measurements because the speech varies so rapidly in intensity. All that can be done is to adopt some definite, even if quite arbitrary method of measurement so that tests taken at different terminals which work together may be comparable.

Operation of a Circuit [11]

It may be of interest to outline the methods of operation and control adopted by the British P.O. at their London terminal.

Each set of terminal equipment is in charge of a technical operator who, in association with his colleagues at the other end, is in control of the channel.

The actual connection of subscribers, etc., is carried out by traffic operators in the Radio Section of the International Telephone Exchange which is situated about eight miles from the terminal. The switchboard equipment and method of operating is very similar to that of an ordinary trunk exchange except that the calls are supervised throughout in order that extra time may be allowed if circuit conditions become bad.

Order wires extend from the technical operator's position to the traffic operator and to the wireless transmitting and receiving stations.

About half an hour before a service is scheduled to be available, the transmitters, etc., are started up and the two technical operators get into touch with each other, the preliminary calling being usually by tone telegraphy. The circuit is then "lined up" by passing a pure tone from the oscillators at each end and the reading of the volume indicator across the outgoing line when the transmitter is fully modulated noted. The noise level is also measured and the singing suppressor adjusted accordingly.

The traffic operators are now notified that the circuit is available by an arrangement of coloured lights which show whether the circuit is fully available or cannot be satisfactorily extended by long line circuits. This latter condition may arise if the signal/noise ratio over the wireless circuit is unusually poor.

The technical operator can light up calling lamps on the traffic operator's position if the distant terminal is calling. This is necessary since the traffic operators are not attached to any particular channel, whilst the technical operators are always on duty on their particular channel.

As different subscribers are connected to the circuit the technical operator will adjust the gain of the transmitting amplifier to keep the wireless transmitter fully modulated at the peaks of speech. This is very important since if the transmitter modulation falls off seriously the noise level at the receiver will become poor.

In recent years attempts have been made to keep the trans-

mitter fully modulated automatically, so that continuous attention by the technical operator is not necessary.

An amplifier giving a constant output for a 50 db range of input is inserted into the line from terminal to transmitter. It is necessary to arrange that the amplifier does not rise to a high gain (in an attempt to produce the constant output) during pauses in speech, even if these are of long duration, as excessive noise would then be introduced and the suppressors would probably lock. In the absence of speech, the amplifier used decreases its gain but only at the rate of 5–10 db per hour. When speech is again applied at the input, however, it quickly resets to the gain necessary to provide the standard output.

Ultra Short-Wave Telephone Circuits

These have come into use in many countries as short links in trunk telephone circuits, where natural barriers make line circuits expensive to install and maintain. Since such links form part of a main trunk line, they must give similar performance as regards stability and frequency-response and must have adequate signal/noise and cross-talk ratios.

They must be designed so that they can operate unattended and require the minimum of maintenance, because the wireless stations will often be sited in high and rather inaccessible places in order to shorten the effective path as much as possible. Provided the propagation distance is commensurate with the optical range, waves within either the metre or centimetre band can be used. Because of the comparative stability of the propagation over such links, they can be operated without the elaborate terminal equipment necessary on the short and long-wave circuits.

In this country, the Post Office, who commenced to use U.S.W. links as far back as 1932, employs them extensively as links between Great Britain and the outlying islands (such as the Channel Islands, the Scillies, the Orkneys and the Shetlands) and also across the short sea path between Scotland and Northern Ireland.

Carrier frequencies between 35 and 80 Mc/s have been used, the transmitter power varying from a watt or two for the shortest links to 100 W on the longest circuit, that between the Orkneys and Shetlands. The effective path length is, of course,

dependent upon the elevation of transmitter and receiver above sea level as well as upon the actual distance. The effective lengths of most circuits are just within or beyond the optical and with the longer circuits the ratio is not much greater than 1·5/1.

The Post Office use a straightforward type of crystal-controlled transmitter and either amplitude or frequency modulation. An anode modulator is used for amplitude modulation and a super-heterodyne receiver, with crystal-controlled oscillator. Directional aerials of the rhombic type are employed for both transmission and reception, these having the great merit of simplicity and strength, so necessary at sites which are often very exposed and windswept.

Many of the Post Office links are worked on a multi-channel basis, employing the carrier-current equipment mentioned on page 607. In some cases this channelling equipment is located at the wireless transmitter and receiver, these being connected to trunk exchanges by 4-wire circuits; in others, the lines connecting the wireless stations are also worked on a carrier basis and the channelling equipment is at the exchanges.

In a number of cases the channels 1 to 12 (see page 609) having carrier frequencies between 12 and 56 kc/s come into the wireless station over a line and are there group-modulated by a 60 kc/s carrier to produce frequencies between 60 and 108 kc/s. These are then used to modulate the transmitter. When amplitude modulation is used, the carrier and both side-bands are transmitted, so that, if a 75 Mc/s carrier is taken as an example, the frequency radiated would be 74,892–74,940, 75,000, 75,060–75,108 kc/s. Spacing the sidebands from the carrier in this way reduces the cross-talk between channels due to non-linearity but the equipment has to be designed to keep this to a minimum.

When the Channel Islands link was restored, after the Occupation, frequency modulation was employed, with a maximum frequency deviation of 300 kc/s, there being six channels. It seems likely that all future Post Office installations will use frequency modulation.

Centimetre-Wave Links

Much work was done during the war on the use of centimetre waves for communication and, in particular, on the use of

favourably-placed relay stations, so that distances much greater than the optical range could be covered in a series of "hops."

Since the technique of producing pulse transmission on centimetric waves was so well developed in connection with radar, it was natural to use pulse transmission on these links. The theory of pulse modulation has already been dealt with and also simple circuits for pulse-duration modulation.

Single-channel pulse modulation does not utilise effectively the frequency band, since this is determined by the duration of the pulse and not by the pulse repetition frequency. The intervals between the pulses of one channel can therefore be partly filled with pulses belonging to other channels and synchronised apparatus at transmitter and receiver will be able to separate out the signals belonging to each channel.

This gives a system of channelling, or multiplex, which works on a time-division instead of a frequency-division basis. Each channel is given the exclusive use of the link for a portion of the time instead of being allotted part of the total frequency-band transmitted.

This form of channelling has the advantage that non-linearity in circuits does not introduce cross-talk because only one channel is passing over the link at any instant. Cross-talk will be produced, however, if the bandwidth (in the receiver, for example) is not adequate, so that the pulses are distorted and prolonged, with the result that the pulses of one channel interfere with those of adjacent channels.

An equipment (the No. 10 Set) working on these lines was developed for the British Army and formed the main link with Montgomery's Headquarters during the advance into Germany.

A wavelength of 6 cm was employed and eight telephone channels were carried. The relay stations were installed at intervals such as 25 to 50 miles, or more, this distance depending upon the availability of elevated sites, and the distance between terminals was sometimes 300 miles. Parabolic reflectors were used for transmission and reception.

The U.S. Army developed a similar equipment and a number of commercial links of this kind have been installed in the U.S.A.

The wide frequency bands that can be obtained with centimetric waves make them particularly suitable for the relaying

of television signals, using amplitude or frequency modulation. This has been done in the U.S.A., and the B.B.C. are employing such a system (of G.E.C. design) to relay television from London to Birmingham, where the signals are broadcast on metre waves.

The Marconi Co. have also conducted tests in which the Alexandra Palace television programme was received and used to frequency-modulate a transmitter working on 510 Mc/s. This transmitter used a horn to form the beam, whilst the receiver, 24 miles away, used a parabolic reflector.

CHAPTER XVII

WIRELESS TELEGRAPH CIRCUITS

LET us first consider telegraphy, as compared with telephony. In the last chapter the conversational nature of telephony was stressed and the fact that ordinary members of the public had to operate the system, once a conversation had begun. Further, the somewhat intimate nature of telephony often means that the people engaged in conversation are well-known to each other and recognition of the voice is an expected concomitant of such communication. This requires a high standard of communication, as not only the intelligence but also the subtle differences between speakers' voices must be conveyed.

The case of telegraphy is very different. A member of the public initiates a message (probably by handing it over the counter) to which he may, or may not, expect a reply. He does not assume that such a telegram will be delivered immediately, neither does he expect the recipient to be presented with a facsimile of his message as evidence of origin. In the case of a foreign cablegram, he may (if he takes any notice of advertising posters) indicate a preference for a particular route. In general, however, neither the sender nor the recipient of the message knows, or cares, how the message was sent, or what difficulties may have been overcome in its transmission. Their main concern is that the message should be accurate and delivered in reasonable time. Any reply that may be called for is virtually a separate communication.

The fact that members of the public do not participate in the actual telegraphic transmission, as they do in telephony, means that such messages can be dealt with in any manner desired, either by personnel skilled in the art and trained to work under very adverse conditions, or they may be sent and received by automatic devices. In the earlier days the distortion of the received signal was such that only the former method was possible and the accuracy depended entirely upon the skill of

the operator, but in many modern systems automatic devices are used throughout and accuracy is dependent upon the skill of the technician and the mechanic. Since messages going out from a telegraph office are not definitely related to messages coming in, a system of simultaneous transmission both ways can be adopted if traffic warrants it.

Telegraph Codes and Signalling Speeds

The land-line form of the Morse code is known to most people and has been illustrated in Chapter III, Fig. 18, where the difference between single-current and double-current was explained.

If single-current is used, then any relays and recorders at the receiver must have a " bias " towards the space position, which the current on mark has to overcome. With double-current, on the other hand, the instrument can be " neutral " and the marking current deflects the moving part in one direction and the spacing current in the other. Relays can be more sensitive in this condition and a change in signal strength (on both mark and space) has much less effect than when the marking current is working against a fixed bias.

In wireless telegraphy the commonest method is to employ single-current Morse, but each dot or dash comprises many cycles of R.F. It will be realised that double-current wireless telegraphy is not possible, using straightforward " on/off " keying, because polarity cannot be distinguished.

Telegraph speeds may be quoted in words per minute (w.p.m.), in which case an average word is taken as comprising five letters. The possible signalling speed of lines and apparatus is mainly governed by the duration of the shortest signal element—the dot in the ordinary Morse code. A unit of signalling speed termed the *baud* (after Baudot, a pioneer telegraph engineer) is therefore frequently used. The speed of signalling would be one baud if the shortest element had a duration of one second.

In the case of Morse, the letters are of very unequal length and somewhat different average figures are used by various telegraph organisations. Cable and Wireless consider one average letter in plain English per second to be equivalent to 9·6 bauds, so that if a circuit can deal with 50 bauds its speed

in words per minute would be 62·5. One Morse numeral per second corresponds to about 16 bauds. It will be seen that, if the signalling currents are regarded as an alternating current of rectangular waveform, then one baud corresponds to half a cycle per second and 50 bauds is 25 c/s. It may be desirable, in some cases, to allow for the transmission of the third harmonic and hence the bandwidth might be about 75 c/s.

On submarine cables a modified form of the Morse code—cable code—is used. The dot and dash are equal in length but are transmitted by currents of opposite polarity, whilst zero current represents a space. The dots and dashes of a letter follow one another without a space between, unlike ordinary Morse, in which there is a space (equal in length to a dot) between each element. When using an ordinary Morse key, the lifting of the key between each element automatically records a space but the sender for cable code has two keys, one for dash and one for dot, and the depression of these one after the other forms "dash-dot" without any space between. Thus the upper line of Fig. 385 shows a cable-code "N" followed by an "M."

Owing to the elements all being the same length and to the elimination of the spaces, the number of bauds for one letter per second is only about 3·7. Since the number of bauds over a submarine cable is usually very limited, this is a great advantage.

Five-unit codes are also frequently used, in which all the characters comprise five elements of equal length, the various characters being different combinations of marks and spaces. Allowing for a space between the letters, these codes clearly require 6 bauds for one character per second, but, in practice, some machines require auxiliary signals, raising the bauds to 8·4 in such cases. In the five-unit codes, figures require no more bauds than letters.

Line and wireless systems have, naturally, influenced each other a great deal in the course of their development and particularly in recent years, with the merging of cable and wireless interests. We propose, therefore, to review very briefly the methods of line telegraphy before considering wireless methods.

Line and Submarine Cable Telegraphy

The earliest systems employed hand sending and reception by operators, using the Morse code. The introduction of the Wheatstone transmitter enabled Morse to be sent at a speed limited only by the line and associated apparatus. The speed can be high on land lines but only moderate on long submarine cables. The transmitter itself can run at over 400 w.p.m. and a modified form has been tried out at 1,200 w.p.m. Although these high speeds are available, the trend of line development has been towards moderate speeds, such as 60 to 80 w.p.m., speeds which are possible on many submarine cables. The reason for this is not far to seek.

Messages have to be put into a suitable form for automatic transmission and printed for delivery at the receiving end. These operations are best carried out on printers of the type-writer class, which work at 60 to 80 w.p.m. The high-speed system will require a number of operators to prepare the tape on keyboard perforators, which tape will then be used for high-speed transmission. At the receiving end the signal currents will be recorded on a tape and then this will be deciphered by a number of operators, who will type the message.

Alternatively, we can reduce the speed of each channel to a suitable value for direct transmission and reception by printers. The possible speed of the line may not be utilised, however, by a single channel and we shall therefore arrange for a number of channels to use the same line (multiplex). Even with multiplex transmission, perforators feeding transmitters are frequently employed, instead of direct transmission, because the signalling speed per channel may still be higher than the operator's typing speed, particularly when handling code messages. The bauds for each channel, multiplied by the number of channels, must not, of course, exceed the signalling speed of the line. On land lines a number of channels will be possible but on long submarine cables the number will always be very limited and, in some cases, only one channel can be used.

In the case of line (including cable) circuits, where propagation conditions remain stable over the whole twenty-four hours, the multiplex method is much more convenient.

Multiplex methods for telegraphy are similar in principle to those already discussed for telephony and fall into the same two

classes—" time-division " and " frequency-division " systems. Time-division telephony is a very recent development but time-division multiplex telegraphy has been in use since the early days of telegraphy.

Time-division multiplex clearly requires synchronism between the sending and receiving apparatus, which therefore runs at a steady speed and is usually controlled by tuning-fork oscillators at each end. Apparatus at the receiving end automatically compares the phase of the received signal with that of the receiving apparatus and corrects any phase shift.

If frequency-multiplex (often called " voice-frequency channelling ") is employed, then, since the frequencies involved in telegraphy are low, a number of channels can be used, all having carrier frequencies within the speech band. Such a multiplex system can therefore be applied to any line or apparatus in place of one telephone channel. The bandwith required by each channel will clearly depend upon its signalling speed. The long submarine cable could not deal with the voice-frequency currents without undue attenuation and this system is limited to land lines and short cables.

Automatic printers are simpler and more satisfactory if they are arranged to run at precisely the same speed as the sending apparatus and hence synchronism between sending and receiving end is usual in modern telegraphy, whether time-division multiplex is being used or not. Automatic printers working off cable code are widely used on submarine cables, whilst on land lines teleprinters using the five-unit code are favoured.

The introduction of synchronism between transmitter and receiver made possible the development of the regenerator. On long-distance telegraph circuits some form of distortion is bound to arise ; in the case of cables or lines such distortion will be due to the long time-constant of the cable or to restricted bandwidth, resulting in the received waveform being badly distorted. It has been the practice for many years to regenerate the received signal by selecting, for a very short time period, a portion of the waveform at a point near its greatest amplitude and using the pulse so selected to control relays or trigger circuits, so producing a distortionless rectangular waveform of the correct duration. Regenerators may be electro-

mechanical or electronic and may be adapted to work on " start-stop " channels—that is, channels using printers which re-start for each character, instead of running continuously.

The possible signalling speed of a submarine cable of a given pattern varies inversely as the square of its length. It has always been the practice, therefore, to land ocean cables on any available islands lying near the route and to set up repeater stations. In the earlier days the signals were read and afterwards retransmitted by operators but the regenerator is now used, which receives the distorted signal from one length of cable and passes on a properly-formed one to the next length.

Operation of Wireless Telegraph Circuits

On wireless-telegraph links, high signalling-speeds are frequently possible but propagation conditions are very variable through the twenty-four hours. In consequence, the Morse, high-speed, single-channel system of operation, previously discussed, has its merits and is widely used. Traffic (particularly cheaper-rate, deferred traffic) can be prepared in advance and then transmitted at high speed when conditions are good. The use of a tape record, read off by skilled operators, also means that more distorted signals can be utilised.

Morse printers (working at moderate speeds) have been developed, but it is usually found better to use a reperforator, that is, a machine which takes in Morse at a reasonably steady speed but not synchronised, and prepares a tape which can then be run through a printer.

Regenerators have been adapted for use on wireless circuits. In this case the received signal is usually of rectangular waveform but the function of the regenerator is to eliminate distortion caused by multi-path effects, since the selection is made at approximately the centre of the time period of each modulation element.

In the Cable and Wireless organisation, messages come into the same central office over both cable and wireless links and it is a great advantage to make the operating methods and apparatus the same for both, as far as possible. Direct connection of a wireless link to a cable link (or vice versa) is also made possible. Hence synchronised, cable-code working, with automatic printing, is widely employed on the wireless links of

Cable and Wireless. Automatic printing has only been made satisfactory by utilising every means for obtaining as steady a signal, free from interference, as possible.

Although earlier wireless systems operated direct from terminal to terminal over the greatest distances possible on earth (England to Australia, for example) Cable and Wireless have nevertheless found it advantageous, on some circuits, to follow cable practice and to use automatic relaying, by regenerator, at judiciously chosen points. By so doing, the reliability of the link and the possible hours of working have been increased. The very important circuits across the North Atlantic, for example, suffer in performance because the great circle path traverses such high latitudes where magnetic storms are troublesome. In this case, transmission via a more southerly relay station will yield a more reliable signal, even though the total transmission distance is considerably increased. Cable and Wireless have, therefore, established three main relay stations at Barbados, Ascension and Colombo.

Cable Code on Wireless Circuits

It will be seen that it is not possible to apply the cable code to a wireless circuit without modification in some way, because

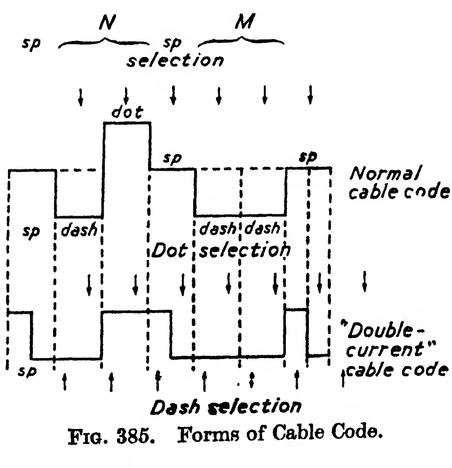


FIG. 385. Forms of Cable Code.

it would not be possible to distinguish between dots and dashes.

One method which may be employed when cable code is used on wireless circuits is illustrated by Fig. 385. The upper line

shows the normal cable code with the points at which the synchronised regenerator would " sample " the signal. The lower line shows a form of double-current cable code, the dot being a marking current and a dash now being signalled by zero current. A space is formed by a current for half a signal length, followed by zero current for the other half.

The selection for dots and dashes now has to be done separately, the dot selector functioning after three-quarters of the signal element duration and the dash selector after one-quarter. The dot selector only responds when there is current, whilst the dash selector only responds to zero current. It is clear that, in the example given, the dash selector will first

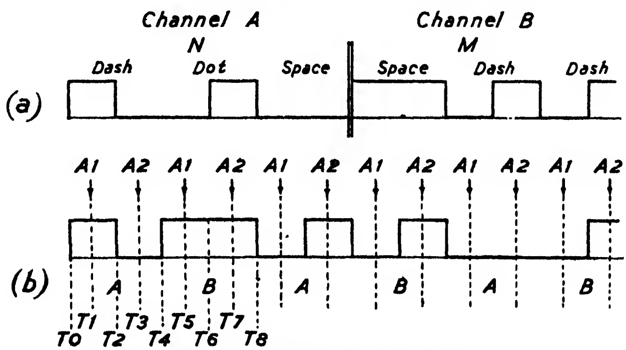


FIG. 386. Modified Cable Code.

function, then the dot, thus recording the letter " N." During the next signal element, neither selector functions and a space is recorded, and so on.

Since this code requires only " on " and " off " for its transmission, it can clearly be applied directly to an ordinary wireless transmitter.

When a two-channel multiplex is being worked over a wireless link, the above method may sometimes lead to long periods without any current. As an extreme case, if each channel made the figure " 0 " at the same time, they would be sent to the line consecutively and there would be no current for ten consecutive signal elements. This may cause trouble with the A.G.C. in the wireless receiver.

A modified arrangement is therefore usually employed. illustrated by Fig. 386. The channel A is shown signalling

"N's" and channel *B* "M's." In the case of *A*, a dash comprises current for the first half element, followed by no current for the second half, whilst for a dot the current occurs in the second half. A space is formed by zero current for the complete element. For *B* these conditions are reversed, a dot

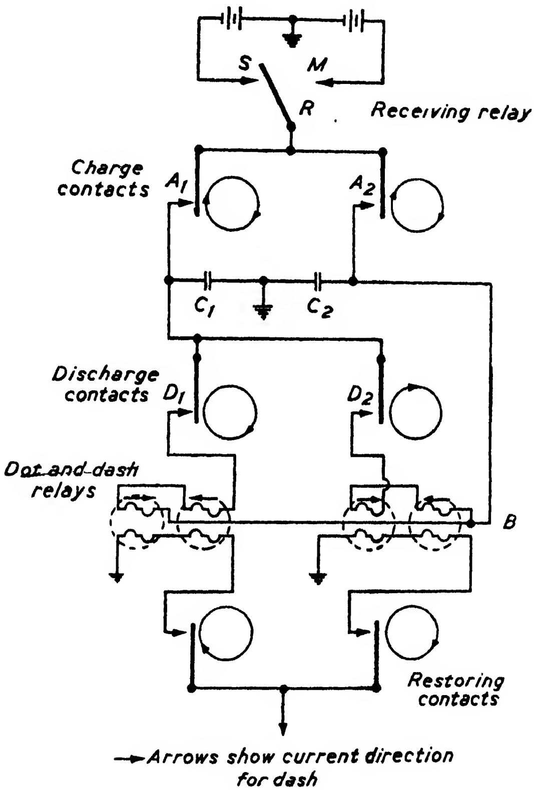


FIG. 387. Circuit for Modified Cable Code.

being the same as *A*'s dash, a dash identical with *A*'s dot and a space comprising current for the whole element.

Fig 386a shows the signalling currents if each channel was working alone. In order to combine the channels, the running speed of the transmitters is halved but the duration of each signal put to line is not increased. The relative phase of the two transmitters is then adjusted so that the two sets of signals are interleaved. The combined signal is, therefore, as in

Fig. 386b. The selection points are marked A_1 and A_2 on the diagram.

Fig. 387 illustrates the circuits used at the receiving position to separate and utilise the incoming signals. The Charge, Discharge and Restoring Contacts are all controlled by cams rotating in synchronism with the signals and the arrows indicate the relative timing of the cams. One half-revolution corresponds to one complete signal element. The Dot and Dash Relays are polarised and remain as set until the Restoring Contact sends a spacing current.

The sequence of operations will best be followed by setting out in tabular form the happenings for each of the instants shown on Fig. 386.

T_0 R on M.
T_1 A_1 charges C_1 +.
T_2 R moves to S.
T_3 A_2 charges C_2 −.
T_4 D_1 discharges $C_1 + C_2$ through the Dot and Dash Relays of Channel A in the dash direction. Dash is recorded until Restoring Contacts send spacing current through relay.
T_4 R moves to M.
T_5 A_1 charges C_1 +.
T_6 No change.
T_7 A_2 charges C_2 +.
T_8 D_2 connects $C_1 + C_2$ through Dot and Dash Relays of Channel B but charge on $C_1 + C_2$ is zero. These relays therefore record space.

And so on.

With this system, the longest period that the combined signal can remain either " on " or " off " is two signal elements. It will be seen that in both the above systems the shortest signalling element corresponds to half a dot or dash. In consequence, the bauds required for a given speed in w.p.m. will be double that for ordinary cable code but this is still less than that required for ordinary Morse.

It is found that errors caused by signal distortion, fading or interference are very rare when using this code with automatic printing. This is because a dot or a dash comprises two signal elements and failure of either results in a space being recorded.

This is usually easily detectable by the operator who scrutinises the message whilst pasting it on to the message form.

When conditions are poor it is possible to arrange that both channels handle the same message but with a delay of about two seconds on the *B* channel. The *A* signals are stored and released at the same time as the *B* signals and both supplied to the recording apparatus. If fading is the main trouble on the circuit, then it can be arranged that if a signal element appears in either channel the recorder functions. On the other hand, if multiple path distortion or interference is bad, it can be arranged that only signal elements appearing on both channels are recorded.

It will not be possible, within the space of this chapter, to describe to any extent the actual apparatus used in telegraphy, and the reader is referred to textbooks and papers devoted to the subject. We will, however, describe in outline the operation of the teleprinter, as similar mechanisms form part of most printing telegraph apparatus.

The Teleprinter

This instrument has come into very extensive use on land lines, in fact it is now the standard method of handling traffic on the inland system of the British Post Office. In appearance it is similar to a typewriter and combines the function of transmitter and receiver. In U.S.A. similar machines are called teletypes.

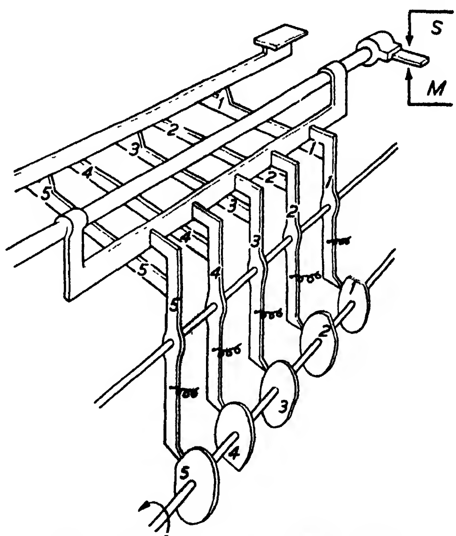


*Fig. 388. Transmitting Mechanism of Teleprinter.

To send a message, the keyboard is manipulated in the usual

* Reproduced, by permission, from Harrison, "Developments in Machine Telegraph Systems"; *J.I.E.E.*, Nov., 1930.

way and at the receiver the message is printed on either a page or tape. A copy can also be made at the transmitter.

A five-unit code is employed. The elements of the mechanical features will be briefly described and these are also typical of the mechanisms used in other printing telegraphs already mentioned.

The principle of the transmitting mechanism is illustrated in Fig. 388. Five " combs " run at right angles to the key levers,

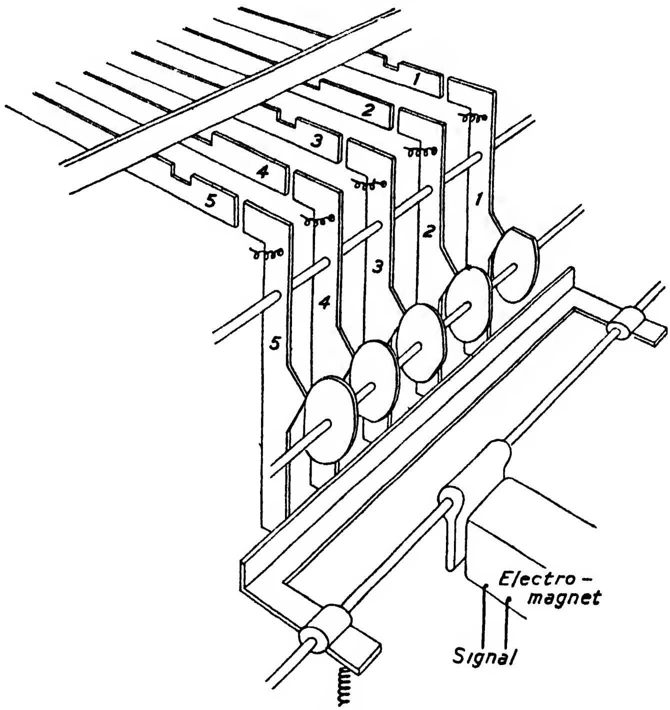


*Fig. 389. Receiving Mechanism of Teleprinter.

only one key being shown on the figure. These bars have projections according to the code for the particular letter key above them. When the particular key shown is depressed, it will be seen that bars 1, 3 and 5 move to the left, whilst 2 and 4 remain stationary. Directly the key is depressed the camshaft is coupled up to the driving motor and commences to revolve counter-clockwise.

* Reproduced, by permission, from Harrison, " Developments in Machine Telegraph Systems " ; *J.I.E.T.*, Nov., 1930.

When the flat of cam 1 reaches the tappet 1, the top of this moves forward, because bar 1 has moved, and puts the contact to mark. When the flat of cam 2 reaches tappet 2, however, the latter cannot move because it is held by bar 2. Hence a space is transmitted. As the cams continue to revolve, 3 makes a mark, 4 a space and 5 a mark. Thus the required code signal is put to line and the cam then stops until another key is depressed.

The principle of the receiving mechanism is shown in Fig. 389. Again there are five bars, usually termed " combs," these having a pattern of slots cut in them. Only the slots required for the same letter as considered in connection with the transmitter are shown in the figure, together with the " latch " for printing that letter.

Directly any key is pressed on the transmitting keyboard a signal is sent to line and this starts the camshaft at the receiver rotating at a speed approximately in synchronism with that in the transmitter. The electro-magnet will be energised from the transmitter at the instant that the flat of receiving cam 1 reaches the tappet, which will be free to move. The top of tappet 1 therefore moves forward and pushes comb 1 along.

When the flat of cam 2 reaches tappet 2, however, the electro-magnet is not energised and tappet 2 is therefore held by its foot and comb 2 is not moved. Thus the result of the signal from the transmitter is that combs 1, 3 and 5 are moved to the left, whilst bars 2 and 4 remain stationary. But it will be seen that the slots for this particular letter have been so placed that moving combs 1, 3 and 5 puts all the slots in line so that the latch shown (and only that latch) can fall into the slots, which movement results in the letter being printed.

When the last unit of the character has left the transmitter an impulse is sent which stops the camshaft at the receiver. Thus seven impulses are transmitted from each character, though only five make up the code. Since the camshafts start for each character, differences in their speeds are not cumulative and elaborate speed control is not necessary.

When the teleprinter is used on land lines the link can be worked on the D.C. double-current system but the instrument is often used to key voice-frequency currents so that telephone lines can be used and also multiplex working.

Teleprinters or teletypes are being increasingly used on long-

distance wireless circuits, but owing to the possibility of undetected errors with the five-unit code they are only successful on the highest-grade wireless circuits.

It is usual to arrange that the teleprinter keys a voice-frequency current of one frequency for mark and another for space, instead of one frequency for mark and no transmission for space, as in line working. The two frequencies may then be used to modulate a radio transmitter or, alternatively, one radio frequency may be transmitted on mark and another on space (frequency-shift keying, to be described later).

The use of two tones results in a double-current system, which we have already seen to be advantageous. The two tones, after being rectified separately, are passed through opposing coils of a relay, which can be "neutral." The risk of printing wrong characters due to fading or distortion, so much more likely to arise on radio links than on lines, is reduced by the use of two tones.

Voice-Frequency Multiplex

One standard teleprinter, voice-frequency multiplex system used on lines in this country, works on frequencies from 420 c/s to 2,460 c/s. The channels are spaced 120 c/s apart, so that 18 channels are accommodated. A simpler system having only four channels is also much used.

Such systems have been successfully applied to wireless working by using two of the adjacent frequencies for each transmission. The channels can therefore be applied to any radio transmitter adjusted for telephony transmission and we have reached the same position with wireless transmission as in line working—the same link may be used for one telephony channel or a number of telegraphy channels. In fact, one side-band of a suppressed-carrier transmitter may be carrying a telephone channel and the other a number of telegraph channels.

In all multiplex working of this kind it will be evident that if there are n channels, then the modulation imposed on the transmitter by any one channel must not exceed $100/n\%$, so that if all the channels come on to mark at the same time the modulation does not exceed 100%. Hence, when conditions are good, multiplex working may be possible but when they are poor it will be necessary to revert to a single channel and

to increase the depth of modulation which this produces at the transmitter.

The R.C.A. Time-Division Multiplex

A form of time-division multiplex specially suited to wireless circuits has been developed by the R.C.A. and is in use on the London-New York circuit. This employs a seven-unit code in which all the characters have four spaces and three marks, the different characters being distinguished by the order of

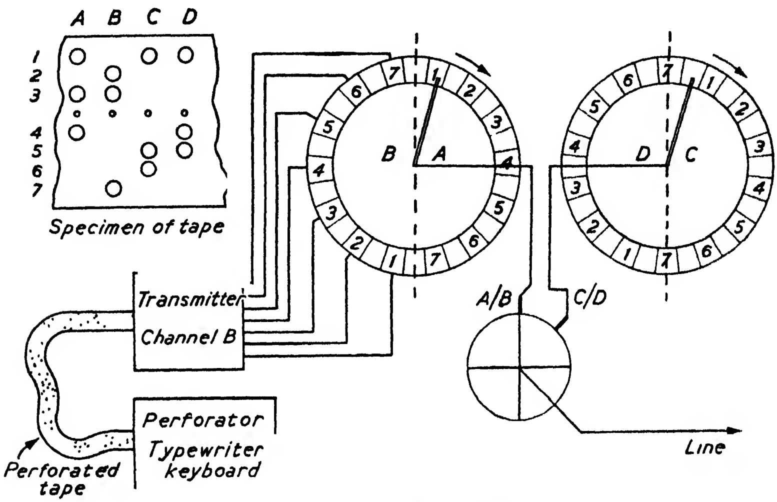


FIG. 390. Time-division Multiplex Apparatus.

spaces and marks. With this code, if one of the marks drops out due to fading, or if a mark appears instead of a space, because of interference or distortion, then the signal does not conform to the code and a star is printed on the tape instead of a wrong character.

The equipment provides for eight channels, each with a speed of 43 bauds (61 w.p.m.), but the aggregate speed of 344 bauds is too high for normal conditions on the transatlantic route and four channels with an aggregate speed of 172 bauds are used.

An outline of the arrangement at the transmitting end is indicated by Fig. 390 in which provision is made for only four channels and the transmitter of only one channel is shown.

The transmitting operator manipulates the typewriter keyboard of the perforator and thereby perforates a specially wide tape. This has seven lines of holes (as well as the driving holes) so that the whole character is on the same transverse line.

It will be seen that there are two distributors of a face-plate type (one for Channels *A* and *B* and the other for Channels *C* and *D*) and also a commutator.

Let us follow the transmission of the letter " A " on Channel *A*. The tape has set up the transmitter so that, when the left-hand distributor arm passes over the A segments, a mark is sent to line on 1, a space on 2, a mark on 3, and so on. After the seventh element, a pulse from the distributor moves the tape forward ready for the transmission of the next character during the next revolution of the distributor. On this channel the mark is made by putting positive to line.

Whilst one distributor has been collecting the signal from *A* in this way, the other has been collecting from *C*, but on *C* channel, mark is made by negative to line. The commutator puts the elements from *A* and *C* to line alternately. Thus there is a sequential (character by character) combination of *A* with *B* and of *C* with *D*, and there is also an interleaving (element by element) of *A*/*B* and *C*/*D*.

The instant of passing a signal element to line is entirely controlled by the commutator, which actually triggers a valve circuit, producing a rectangular waveform to line.

An identical arrangement is used at the receiving end for distributing the signals to the automatic printers of each channel.

The face-plate distributors and the commutator at each end are driven by a synchronous motor, the standard of frequency being a 600 c/s tuning fork (valve-maintained). This is very stable in frequency, but, in order to correct for any slow drift, there is an arrangement whereby, if the receiving distributors are not in phase with the incoming signals, then a relay switches on a small D.C. motor which drives (through gearing) the frame of the synchronous motor and thereby corrects the phase.

Recording of High-Speed Signals

It would appear to be a simple matter to rectify (and if necessary amplify further) the beat frequency output of a

receiver taking C.W. telegraphy and to supply the resulting D.C. signals to a recorder such as the undulator mentioned previously. The matter would be simple if the output of the receiver consisted of constant amplitude, undistorted code, but this will rarely be the case.

In Chapter V it was pointed out that a long-distance, short-wave signal is often received by two or more paths, having delay times differing by a milli-second or so. At 125 words per minute the length of a dot is 10 milli-sec. and the multiple reception may increase this length at the receiver by 2 milli-sec.

Any form of distortion which causes either the beginning or end of the signal to shift from its correct time is termed phase distortion and it is very variable on most short-wave circuits.

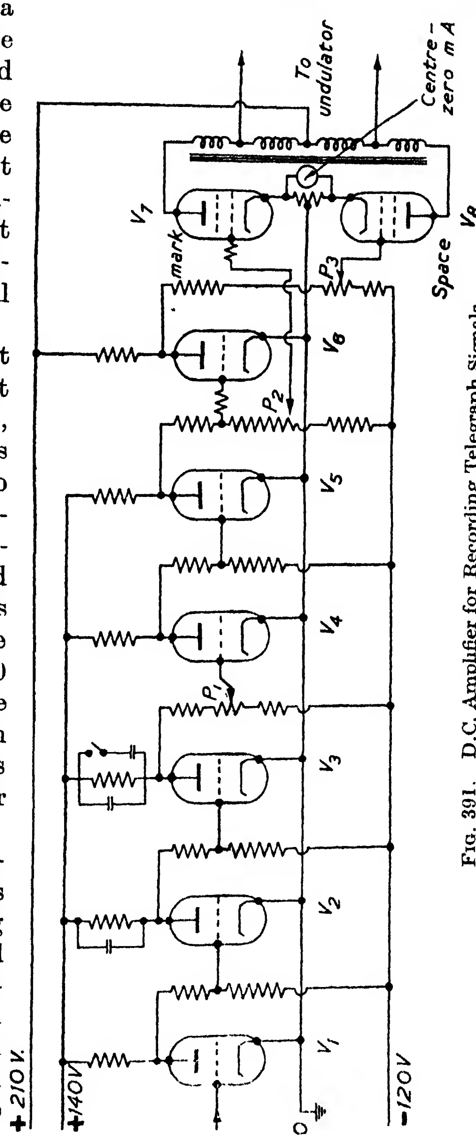


FIG. 391. D.C. Amplifier for Recording Telegraph Signals.

In addition, of course, there will usually be considerable fading, which may vary the signal amplitude as much as 20 db

and have a frequency which is greater than the keying speed. The A.G.C. on the receiver will result in an output which is comparatively constant but further levelling out will usually be necessary in the recording circuits.

The circuits for rectifying the V.F. output from the receiver have already been discussed in Chapter XV, in order to explain diversity methods. We shall now consider suitable circuits to follow the detecting stage and will base our discussion on the Marconi RB150 equipment.

We have now to deal with D.C. amplification, because the recorder must remain in the correct position if there is either a long space in the transmission, or a long dash transmitted. The form of coupling will be clear from Fig. 391.

The operating conditions of the valves are summarised below :—

Valve	Anode Current	
	Space	Mark
V_1 . . .	Full	Reduced
V_2 . . .	0	Some
V_3, V_5, V_7 .	Full	0
V_4, V_6, V_8 .	0	Full

The condensers in the anode circuits of V_2 and V_3 remove any V.F. ripple remaining from the rectifier circuit but the effect of these condensers is also to convert rectangular input pulses into approximately triangular ones, due to the exponential charge and discharge curves of the condenser-resistance combination.

The potentiometer P_1 (the Signal Bias control) enables any desired width of this signal to be selected and hence "neutral" tape (that is, a record in which dots and spaces are of equal length) to be obtained.

V_5 is a second limiter which squares the waveform selected by V_4. V_6 is a phase-splitting valve, to provide a voltage on the grid of V_8 which is opposite in polarity to that supplied to V_7.

Since the undulator will usually have a low impedance, a step-down transformer is used in the output but this must be an auto-transformer in order that there may be a current through the undulator during a long mark or space.

Relaying Signals to a Central Telegraph Office

We have so far considered the recorder to be adjacent to the receiver. A monitoring recorder will usually be available at the receiving station but the signals will normally be transmitted over a line to a central office.

Telephone lines will generally be available and voice-frequency signalling is the most convenient method of transmission. Where a large number of channels are in use voice-frequency channelling is used.

The standard channelling equipment mentioned on page 664 is not normally suitable unless the radio channels use tele-

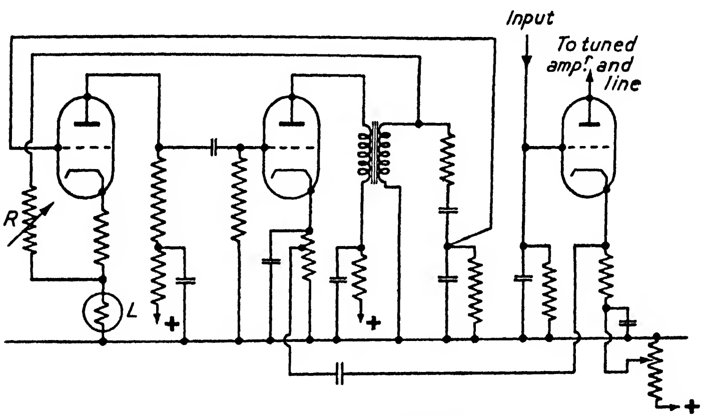


Fig. 392. Tone Sender.

printers or otherwise operate at a slow speed, and it is usual for communication organisations to provide special equipment. In addition to the fact that most radio circuits are operated at speeds between 80 and 160 bauds the speed requirements are also increased due to distortion ; for instance the addition of 2 milli-sec. multipath distortion to a modulation element at 100 bauds will cause a momentary increase of speed to 125 bauds. Unless the line channels are designed for this additional speed margin it is possible for a complete failure of an otherwise readable signal to occur at the central office.

Channels are, therefore, often designed for a maximum of 120 bauds for normal speed circuits and for a maximum of 240 bauds for the higher speed circuits. In some organisations

both normal and high speed channels may be provided on the same line. In the Cable and Wireless system, normal-speed channels have a spacing of 240 c/s and at high-speed a spacing of 480 c/s. Twelve 240 c/s–spaced or six 480 c/s-spaced channels can be accommodated on one line.

On smaller stations where only a few channels are required, the tone-sender forming the transmitting element of the V.F. channel may be incorporated in the radio receiver, but normally the channelling equipment will be grouped centrally and the channel transmitting units controlled by D.C. potentials from the radio receivers.

A typical tone-sender for incorporation in the radio receiver is the type RT150, which employs a resistance-capacitance oscillator, the basic circuit of which is shown in Fig. 392. A two-stage, resistance-capacitance amplifier is coupled back upon itself and oscillation is therefore produced at a frequency for which the total phase shift round the circuit is 360°. The resistance-capacitance network is actually switched, so that six, spot frequencies between 600 and 1,800 c/s can be obtained.

Negative feed-back is arranged through the resistance R, the amount of feed-back being evidently dependent upon the ratio R/L, where L is the resistance of the lamp in the cathode circuit. If the negative feed-back is set to a value which permits a small oscillation of good waveform and then the oscillation tends to increase, the increased current through L will raise its temperature and hence its resistance. This will increase the feed-back and prevent the oscillation building up.

The output of the oscillator is connected (by cathode-coupling) with an amplifier, to the grid of which the D.C. output of the recording circuit is connected. When the recorder is on " space," the grid is made so negative that no tone signal passes through this valve but when the recorder is on " mark " the bias is reduced and tone passes to line.

At the central office a recording unit similar to that already described, but somewhat simpler, will be necessary.

Similar V.F. equipment is used for the control of the radio transmitters from the C.T.O. except that in this case the tone sender is controlled from the telegraph transmitting equipment and the recording unit is at the transmitting station and controls the wireless transmitter.

In some modern equipments, frequency-shift keying (see next section) is used on the V.F. channels between wireless stations and C.T.O., even when the wireless circuit is working with "on–off" keying, with a consequent reduction of distortion.

Frequency-Shift Keying

An old method of signalling has recently been revived, as it offers certain advantages. When the arc was employed as a transmitter, it was not possible to key this by "on–off" methods because the arc would not restart. Signalling was therefore carried out by changing the frequency sufficiently between "mark" and "space" to make a large difference in the heterodyne beat note, or even to make the signal inaudible during space.

In the modern method, used for high-speed recording on short waves, the difference in the transmitter frequency between mark and space is usually 850 c/s, or 500 c/s, and it will be realised that, if the circuit is going to work without undue attention, the receiver oscillators must be very stable and some form of automatic frequency control will be desirable.

The method will be discussed in relation to an equipment recently designed, using the CR150 receiver already described. A block schematic is shown in Fig. 393, from which it will be seen that two receivers are used for double-diversity reception and the A.G.C. circuits are combined in the way discussed previously.

In the F/S equipment, combined oscillators are used for all three frequency changers and an automatic frequency-control is fitted to the second oscillator. This is worked off the mark signal and ensures that this has a frequency of 2,125 c/s. The signal then passes through a filter (which accepts both mark and space but reduces noise at frequencies outside its pass-band), followed by a limiter, of a type similar to that described on page 575 in connection with the reception of F.M.

The mark and space frequencies are now separated by filters and separately rectified and amplified. The resulting current from mark is fed to one coil of the polarised relay and the space current to the other. The output of receiver *B* has passed through identical stages and is now also fed into the same relay coils.

The combined A.G.C. has ensured that the receiver with the strongest signal at any time contributes most of the output but the noise suppression circuits also assist in seeing that the signal with the best signal/noise ratio contributes most current to the relay.

A frequency-indicator is fitted which shows on a meter the actual difference between mark and space frequencies, provided that these are not more than 100 c/s from their correct value of 850 c/s.

In the method shown above, each receiver is, in effect,

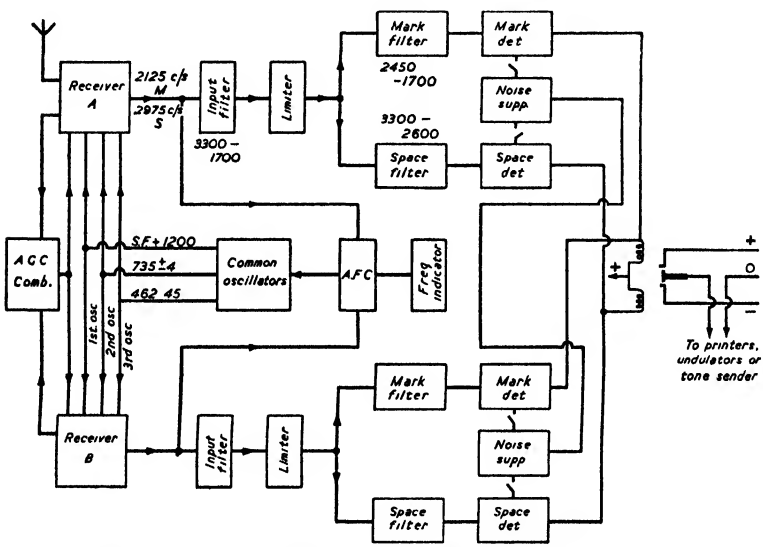


FIG. 393. Schematic of Receiver for Frequency-shift Keying.

handling two " on-off " signals, the mark and space detectors functioning in the ordinary way. An alternative is to use F.M. circuits, that is, to pass the complete signal through a limiter followed by a discriminator.

Extended tests have shown that frequency-shift keying can usually yield better results than " on-off " keying. It is seen to be a double-current system but the difference in frequency between mark and space may mean that selective fading causes them to be quite different in strength at the receiver. From this point of view it would seem that the frequency shift should be small—say 100 c/s. Difficulties would arise, however, from

frequency instability and the design of filters would be difficult if the receiving circuit was of the type shown in Fig. 393. Many tests have been carried out with a 500 c/s shift.

The fact that energy is reaching the receiver during space, as well as mark, means that a transmitter of given peak power can deliver twice the energy to the receiver. Limiting can be very effective (as with F.M. reception) but good limiting, after detection, is also possible when receiving " on-off " signals.

Picture Telegraphy

Facilities now exist for the transmission of pictures over many wireless circuits and there are a number of different

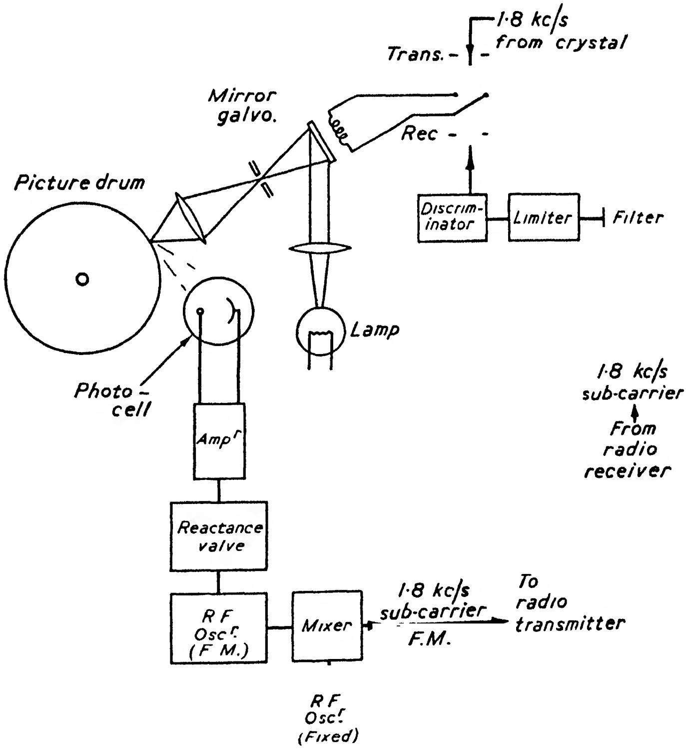


FIG. 394. Facsimile Equipment.

designs of apparatus. All existing systems use reflected light from the picture to be transmitted. The picture may be

recorded at the receiver on film, light sensitive paper, or various chemical processes, but really good reproduction can only be obtained by photographic methods.

To convert the varying reflected light into a varying electrical current, the photo-electric cell is used in nearly all systems. A variety of methods have been used, however, to produce a varying light at the receiver.

As it is not possible to describe all the present-day apparatus giving good results on wireless circuits, we propose to describe only that built by the General Electric Co. to Post Office specification for Cable and Wireless.

It is shown in schematic form in Fig. 394, the same drum and mirror galvanometer being used for transmission and reception. The picture to be transmitted is wrapped round the drum and this revolves on a lead screw at 60 r.p.m. and at the same time travels slowly along so that the whole picture is scanned slowly by the spot of light in about ten minutes. The synchronous motor driving the drum is supplied from a 108 kc/s crystal oscillator, the frequency of which is first divided by 10 and then by 6 to give 1·8 kc/s.

If a steady light was shone on to the picture, then the photocell current would consist of steady current modulated in accordance with the picture and a D.C. amplifier would be necessary. This is difficult to design where high gain from a small input is desired. The light is therefore interrupted at 1·8 kc/s by the mirror galvanometer (having a very small mirror and coil) and slit.

The photocell output therefore comprises a 1·8 kc/s carrier, amplitude modulated by the light and shade of the picture. After amplification, this is rectified and applied to a reactance valve (see page 543) to vary the frequency of a radio-frequency oscillator. This oscillator beats with a fixed R.F. oscillator so that when there is no output from the photocell (that is, the picture is black) the beat frequency is 2·3 kc/s, whilst for white the frequency becomes 1·5 kc/s. Thus we have now what is usually termed a sub-carrier of 1·9 kc/s, frequency modulated with a maximum deviation of 0·4 kc/s, the modulation frequency varying with the detail in the picture but rising to a maximum of about 0·8 kc/s.

This sub-carrier can now be sent over a telephone line to a

wireless transmitting station and applied as amplitude modulation to a transmitter.

On reception, the wireless receiver will deliver the sub-carrier frequency modulated over a telephone line to the picture apparatus. Here the circuits are basically those of a F.M. receiver modified because the carrier frequency is only 1·9 kc/s. After passing through a limiter, the carrier beats with 10·8 kc/s derived from the crystal. The lower sideband is then removed and the 12·7 kc/s carrier with its 0·4 kc/s deviation is applied to a discriminator and amplitude modulation produced.

After rectification this is passed through the mirror galvanometer and the same optical system is used in reception, thereby directing a varying light on to the film in the drum.

The crystals remain sufficiently constant in frequency that no synchronising signals are required.

If amplitude modulation were employed the shades of the picture would be dependent upon the fading of the signal and results would be very variable. Frequency modulation gets over this difficulty besides having a better signal/noise ratio.

Ultra-Short Wave Circuits

If approximately optical ranges are used, signals are sufficiently steady and the signal/noise ratio sufficiently good, so that line methods of working can be employed without modification. The high carrier frequency and the short range makes possible the use of high modulation frequencies and hence multiplex telegraphy, with a very large number of channels, can be used.

Ultra-short wave circuits using pulse multiplex for telephony have already been discussed. Similar methods can evidently be employed for telegraphy. In the centimetre-wave circuit mentioned in the last chapter, each of the eight pulse-multiplex telephone channels could accommodate, if desired, an 18-channel, voice-frequency multiplex, making a possible 144 teleprinter channels in all.

Selected References

(1) Mallett. *Telegraphy and Telephony.* Chapman and Hall.
(2) Herbert. *Telegraphy.* Pitman.
(3) Cohen. *A Handbook of Telecommunications.* Pitman.

CHAPTER XVIII

SOME TYPICAL TRANSMITTERS AND A TRANSMITTER-RECEIVER

It is considered that it may be of some interest to have descriptions of actual equipment, in order to see how the various circuits discussed in earlier chapters are fitted together to form a complete transmitter.

Space is not available for lengthy descriptions, even if these were desirable in a book of this kind, and hence our descriptions will be limited to the more interesting features of a general-purpose transmitter, a large broadcast transmitter and a frequency-modulated broadcast transmitter. The chapter also includes a description of a F.M. trans-receiver intended for mobile communication and a brief description of frequency-channel selection.

Marconi Transmitter TFS 31

The medium-power telegraph-telephone transmitter to be described has been designed for communication channels where a number of frequencies may be needed at different times and where rapid frequency-changing is required. Such conditions may arise, for instance, at aircraft ground stations or shore stations for ship to shore traffic. The transmitter, which has a frequency range from 3·0 to 22·2 Mc/s, gives an output of 4 to 5 kW on telegraphy, may be keyed up to 200 w.p.m. by means of a partial absorber system, and is provided with a phase-modulator unit for combating selective fading.

On telephony a carrier power of 3 to 3·5 kW is obtained, high-power modulation up to 100% being possible. The fidelity is such that at 80% modulation the tolerance is within ± 1 db over the audio range of 150 to 5,000 c/s, with a scintillation of less than ± 1 c/s.

A block schematic circuit of the transmitter is shown in Fig. 395 from which it is seen that the R.F. circuit arrangement comprises alternative forms of M.O., a harmonic generator,

and power amplifier, a general view of the R.F. unit being shown in Fig. 396 with the harmonic generator partly withdrawn. The provision of a Franklin M.O. (having a stability of 1 part in 20,000), enables any frequency within the band to be provided quickly, and a crystal unit housing ten separate crystals (with a stability of 1 part in 100,000), enables that number of inharmonically related spot frequencies to be available. Actually the number of spot waves that can be

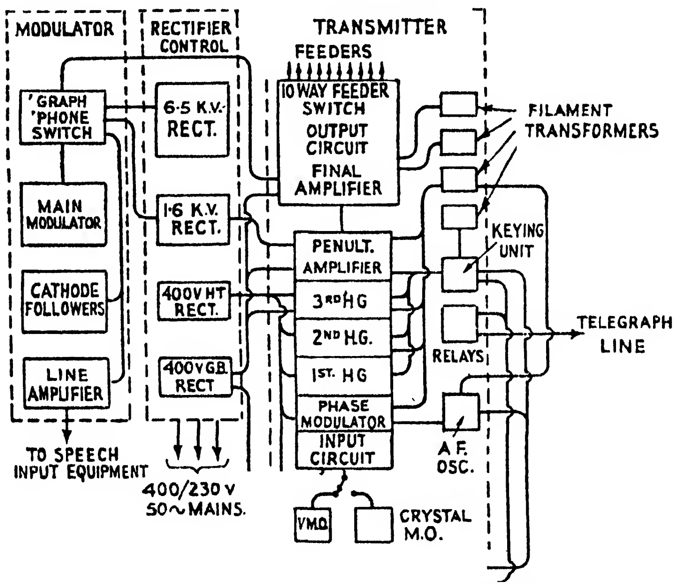


FIG. 395. Block Schematic of TFS31 Transmitter.

provided is not so much a function of the number of separate M.O. frequencies but rather of the mechanical design of the quick wave-changing control. This functions by means of specially-designed, barrel switches correlated with condenser controls, and provision is made on these barrels for ten separate sets of contacts each set of which is associated with a given spot wave, previously set up.

By such a design, to change the frequency involves but the re-setting of six rotary switches fitted with click-stop mechanisms and the readjustment of six continuously-variable controls to predetermined positions, the total necessary actions taking

one man less than ninety seconds to perform. This system also lends itself to wave-changing by remote control, it being

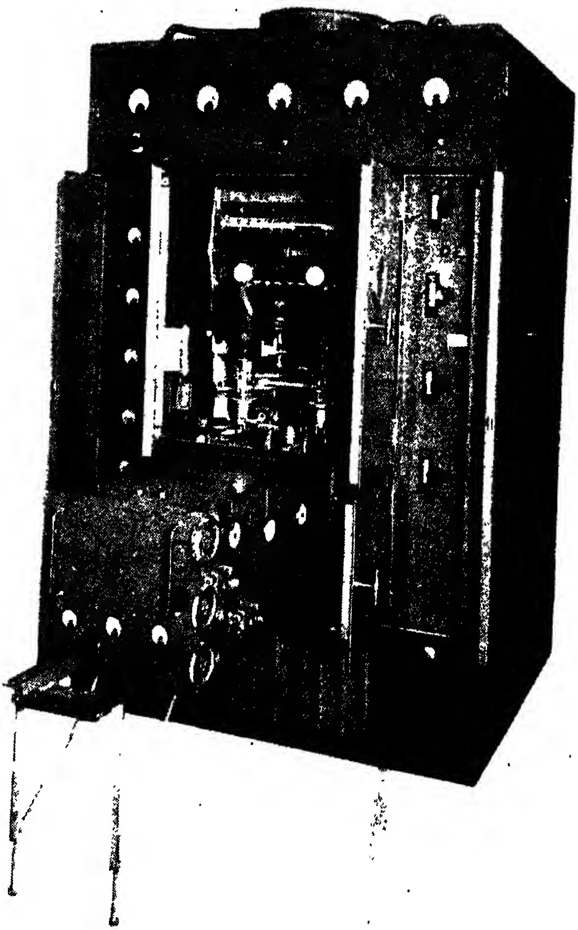


FIG. 396. TFS31 Transmitter.

necessary only to incorporate additional units at each end operated through three pairs of lines.

The output circuit from the power amplifier is designed to

match either a twin-wire or concentric transmission line, and in fact is so flexible that it can handle load resistance values between 40Ω and 1,000Ω and powers from one-quarter to full load without affecting the tuning adjustments of the transmitter. The usual interlocking switchgear protects personnel and ensures correct operating sequence when starting up and shutting down, thus protecting component parts from damage. We will now describe briefly the various circuits.

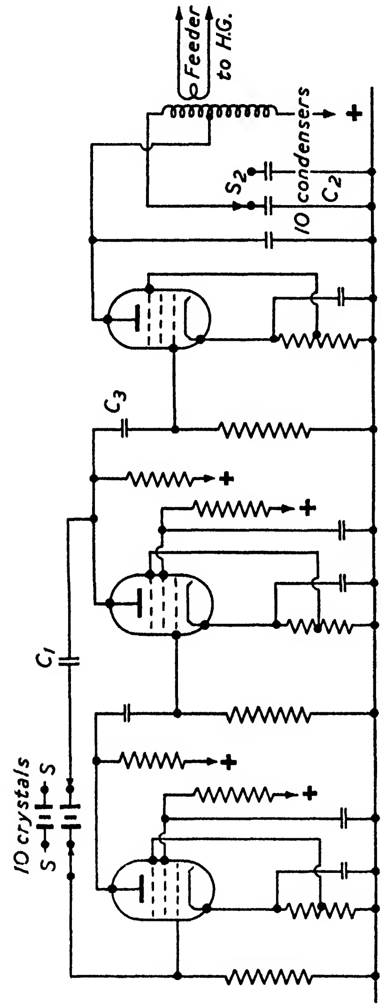


FIG. 397. TFS31 Crystal M.O. Unit.

Crystal M.O. Unit

Fig. 397 shows a simplified circuit diagram of the Crystal M.O. unit, operated from a self-contained power pack. The crystal-maintaining circuit is made up of the two pentode valves, V_1 and V_2, arranged as a resistance-capacitance coupled amplifier, the controlling crystal acting as a feed-back between the output of V_2 and the input of V_1.

Ten A.T.-cut crystals are provided, each in a plug-in holder to facilitate rapid changing, and they are all housed within a thermostatically-controlled chamber.

The oscillator output is applied through C_3 to the grid of the pentode V_3, which acts as a frequency multiplier, the anode circuit being tuned to five times the frequency of the crystal

chosen. The correct pre-set condenser, C_2, to do this is selected by a switch S_2 which is ganged with the crystal-selector switch. Crystals having frequencies between 200 and 266·6 kc/s are used, so that the outputs from the M.O. unit lie between 1 and 1·333 Mc/s, the output coupling being arranged to match an 80Ω concentric transmission line which feeds the harmonic generator.

As mentioned previously, a Franklin M.O. is provided as an alternative to the crystals. This type of oscillator has already been discussed on page 430.

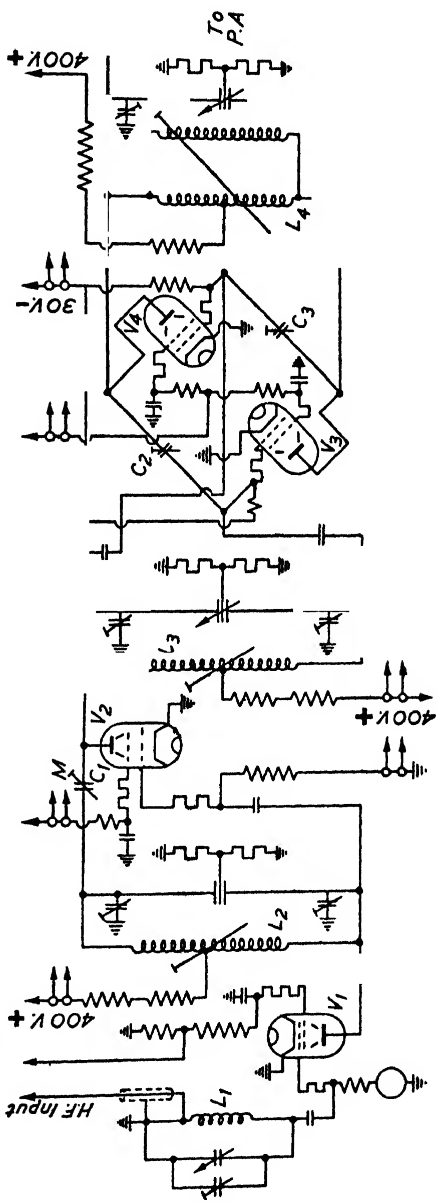


FIG. 39 TFS31 Harmonic Generator.

Harmonic Generator

Since the frequency range of the M.O. unit lies between 1·0 and 1·333 Mc/s, and the transmitter output frequency range is from 3·0 to 22·2 Mc/s, harmonic generation having multiplication as great as 18/1, or as low as 3/1, will be needed, and at all frequencies there must be sufficient gain to drive the power amplifier.

Fig. 398 shows a diagram of connections of the harmonic generator, which has an input circuit, L_1C, tuned to the M.O. (to which is coupled the phase modulator), followed by three frequency-multiplying stages. It will be seen that each stage employs beam tetrode valves, the first two stages employing a single valve and the last two push-pull valves in a bridge network. It will be noted from the figure that the second and third stage are neutralised, the second by the condenser C_1 and the third by the bridge. This is necessary as each may at

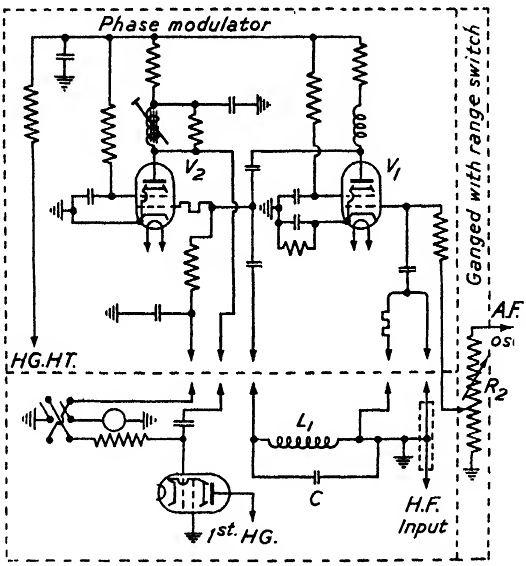


FIG. 399. TFS31 Phase-modulator Unit.

times be functioning as a straight amplifier stage and the valve grid/anode capacity, although small, may be sufficient to endanger stability.

We showed in a previous chapter that for efficiency the inductance of amplifier stages should be as high as possible. With the present transmitter the inductances of each stage have tappings connected to a barrel switch which selects the appropriate maximum inductance, final tuning being by means of small variable condensers. Both barrel switches and condensers are ganged together with a switch which varies the screen potential of the valves so as to vary the gain of the stages.

This additional control is necessary as the harmonic generator has to deliver an approximately constant output on all ranges and the frequency multiplication varies to such wide limits as 3/1 and 18/1.

The Phase Modulator

This unit operates on the input tuned circuit of the harmonic amplifier, L_1C. The reactance valve used is a tetrode, V_1 (Fig. 399), which is connected across the input circuit with grid at the low potential end, the non-inductive resistance in series giving the necessary phase correction of the H.F. potentials which are in phase quadrature on grid and anode. The audio-frequency modulating voltage applied through the potentiometer R_2 varies the grid bias of the modulator, varying its reactance and hence the tuning of the input circuit. In order to maintain constant phase-modulation for the different frequency-multiplication used, the audio-frequency voltage applied is adjusted to be in the inverse ratio of frequency multiplication, and this is accomplished by variation of the potential from R_2, through a barrel switch which is ganged with the harmonic-generator controls. Following the reactance valve is a tuned amplifier, V_2, with heavily damped secondary to restore the loss caused by the reactance valve. The parallel damping is used to ensure approximately constant output over the band.

Power Amplifier

The power amplifier has two stages, the first using two TT10, 100-watt, beam-tetrode valves in a balanced bridge and the final stage two ACT9, air-cooled triodes, again in a bridge balance, the diagram of connections being shown in Fig. 400. In both stages similar methods of tuning are adopted to those for the harmonic generator, namely by tapped or variable inductances adjusted through tappings on barrel switches in conjunction with small variable condensers. It will be observed, however, that two parallel coils are used, as with the final stage of the harmonic generator. The use of two coils, one of which can be used for the lower frequencies, and both in parallel for the higher frequencies, has advantages in that the current carrying capacity is automatically increased at the

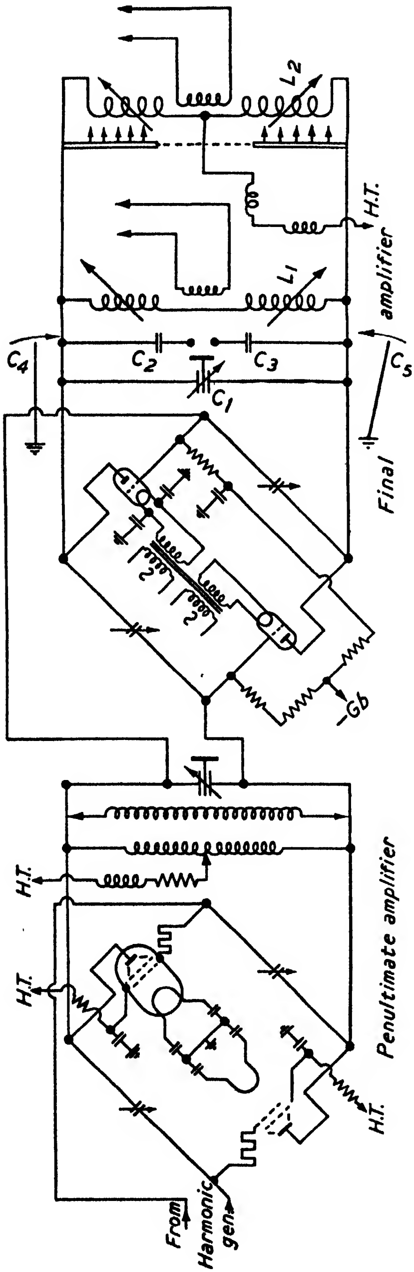


FIG. 400. TFS31 Power Amplifier.

higher frequencies and it solves a difficult problem in providing the overlaps needed when only small tuning capacities are used, since with two coils much greater inductance selection can be obtained, and a more flexible arrangement of coupling to various values of load resistances. Dealing with the final stage particularly, this has two coils, L_1 and L_2, the first and smaller consisting of 12 separate turns which can be set up in different series/parallel arrangements; and the second and much larger coil, a single layer coil tapped so that different turns can be switched into circuit. The coil L_2 alone covers the lower frequency range, i.e. from 5·4 to 3·0 Mc/s, and for the higher frequencies both coils are used in parallel.

The different series-parallel arrangements for coil L_1 are selected by a cage form of switch built of low loss ceramic material to which are secured copper blades

contacting female spring-like jaw connections on the coil, the contacts entering edgewise with a cutting action. A similar design of barrel and switch contacts are used for selecting the number of turns in use, the barrels being ganged to a single handwheel, final tuning being by means of the small variable capacity C_1, to which can be added by an interlocking switch the capacities C_2, C_3.

The output circuit (not shown) is tuned and can be coupled to either the smaller or larger inductance in the anode circuit, the two coupling coils being geared together and rotated simultaneously by a single control handle.

The bridge circuit used is conventional in design, the two condenser plates C_4 and C_5 being employed to equalise the feeds in the two valves. These are merely plates coupled to the fixed plates of the tuning condenser, which as is usual form a main support for the cradles of the valves and about which the circuit is mechanically and electrically built to shorten leads and obtain symmetry for the bridge circuit.

Keying Circuit

Telegraph keying is effected by a partial absorber circuit of the type described in Chapter XIII, page 535. With the present transmitter, the partial absorber is arranged so that on space a negative voltage of the appropriate value is applied both to the screens of the beam tetrode valves in the harmonic amplifier and in the penultimate power amplifier. The method by which the partial absorber works has already been fully described and can be controlled either by relay, or D.C. amplifier operating from a tone-keyed line. With the TFS transmitter, the standard circuit employs a " Creed " relay although alternative electronic control can be used without change of main circuits.

Telephone Modulator

The method of using anode modulation has been discussed already in Chapter XIII. In the present transmitter, anode modulation is applied to the final amplifier and the modulator unit employs three stages, all push-pull, The first two stages employ beam tetrode valves, resistance-capacitance coupled, the output from these driving a cathode-follower stage (see page 521) employing two triode valves with the load, of course, in the

cathode circuit. This cathode-follower gives a small voltage loss, but as it provides considerable power gain the succeeding main modulator can be driven heavily into grid current with a minimum of distortion. The main modulator employs two triodes in Class B, push-pull, and the output is coupled to the H.F. anode circuit through a transformer and modulator choke.

Marconi 100 kW Short Wave Broadcast Transmitter

It may be of interest to consider the arrangement of a high-power, high-fidelity transmitter, with special reference to the final stages. As an example of the specification to which a transmitter of this class has to be designed, the main points concerning the R.F. circuits are given below.

The unmodulated carrier power output to be 100 kW from 80 m to 30 m, falling off to 75 kW at 13 m. A quick change to any one of four pre-set frequencies to be possible. Crystal oscillators maintaining the frequency to within $\pm$ 1 in 100,000 to be provided for each of eleven specified frequencies. In addition, a resonant-circuit oscillator to be provided, to enable other frequencies to be obtained. The frequency stability of this to be within $\pm$ 1 in 25,000. The scintillation during modulation to be less than 1 c/s when the modulation is 80%.

The R.M.S. sum of the harmonic content on any frequency of modulation is not to exceed 4% of the voltage at the fundamental frequency, when the depth of modulation is 90%. The frequency response to be level to within $\pm$ 2 db. from 50 to 8,000 c/s.

Oscillators and Harmonic Amplifiers

There are eleven crystals, all housed in one temperature-controlled chamber. Five independent harmonic amplifiers are provided, so that adjustments for five different frequencies can be set up and a quick change made between them.

Each harmonic amplifier consists of three pentode stages used for frequency-multiplying, followed by three triode amplifying stages, the output from the third stage being about 120 kW, at the final frequency to be radiated.

The crystal oscillator unit and complete harmonic amplifier unit are supplied in duplicate.

Intermediate and Final Amplifier

It would obviously be quite uneconomic to duplicate the high-power stages in the same manner and a normal switching system does not give a sufficiently compact lay-out where large powers are concerned, owing to the length of connecting leads involved. Previously adjusted "plug-in" coils could be used but they are an inconvenient and inelegant solution, and the

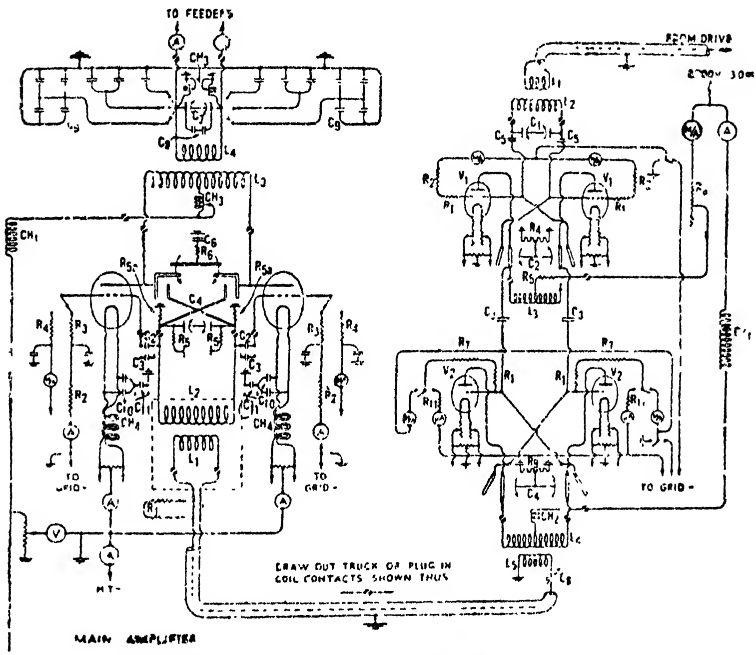


FIG. 401. Amplifier Stages of 100 kW Transmitter.

adoption of a turntable limits the number of circuits that can be pre-set. The solution in the transmitter being described is an ingenious one. A railway track is provided behind the transmitter and the anode and grid inductances are mounted on trucks together with the complete coupling circuit to the feeder. As many sets of circuits as are considered necessary can therefore be provided, and the method has the advantage that none of the idle circuits remain in the high frequency field of the transmitter, and the circuits not in use can be re-set to

any new frequency whilst the transmitter is working. When the appropriate truck is run up into the transmitter, the track ensures exact alignment of the contact system which is a special design of a vice-grip type ensuring a very low-resistance contact. This arrangement of circuit trucks is employed for the last three stages of the transmitter, but whereas the final stage has a truck to itself, the two previous stages have circuits which are mounted one above the other and thus share a truck. A view of one of the trucks is shown in Fig. 403.

An unusual mechanical feature of design is that the movement of variable condensers, feeder coupling, etc., are controlled by a hydraulic ram, this giving a very smooth control and making the position of the various components quite independent of the control-layout of the front-panel.

The amplifier circuits (Fig. 401) are conventional triode balanced bridges as discussed on page 381. Thus considering the final stage L_1, L_2, tuned by its condenser, is the grid input stage; L_3, the output circuit inductance tuned by its condenser, the bridge being balanced by condensers C_4. The grid reactance condensers (the purpose of which was discussed on page 382) are shown as C_2 and the filament reactance chokes and condensers are indicated by CH_4, C_{10} and C_{11}.

As mentioned in a previous chapter, the stability and efficiency of a short-wave transmitter depends, not so much upon the electrical circuit adopted, as upon the mechanical lay-out. Thus, the valves and circuits together must form a symmetrical group to earth, stray capacities must be reduced to a minimum, connecting leads cut to a negligible length, and coupling between the grid and anode circuits eliminated. The anode-circuit tuning inductance must be kept as high as possible and therefore the circuit tuning-condenser as small as possible and all conductor surfaces at high potential must be sufficiently rounded to avoid brushing or torch discharges. Insulators must, wherever possible, be kept outside the high-frequency field, and the dimensions of all conductors must be sufficient to carry the large currents that will be obtained on short waves at high powers. Thus with the present set, the main amplifier H.F. circuit at 13·5 metres will have a peak voltage of 20,000 when delivering full unmodulated output, and an R.M.S. current of 240 amps., rising to 300 amps. on

100% modulation. The grid H.F. current on this wavelength is of the order of 120 amps.

The mechanical lay-out of the main amplifier is shown in Figs. 402 and 403 which illustrates a number of the points mentioned. The circuit is built around two water-cooled

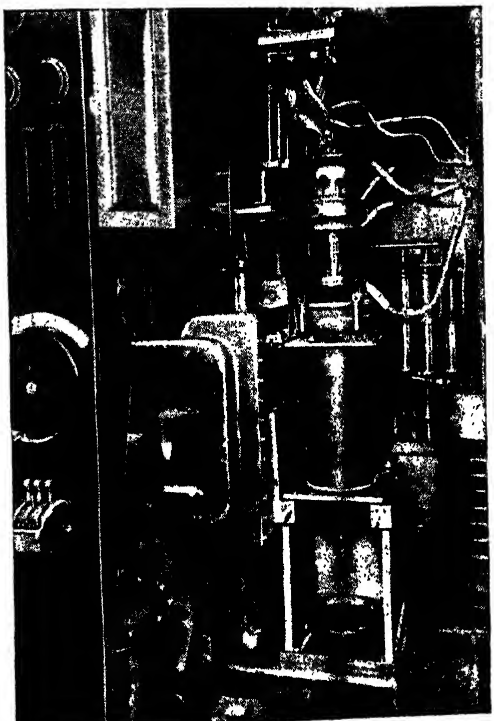


FIG. 402. Lay-out of Main Amplifier.

triode valves of the CAT17 type and from Fig. 402 it is observed that the valve jacket is bolted direct to a plate which serves both as one plate of the tuning condenser and as one plate of the balancing condenser, the other plates of those condensers being indicated in the figure. The thickness of these condenser plates is to be noted, the object being to make possible a large radius of curvature at the edges, thus avoiding corona discharge.

In spite of their apparent massiveness the actual capacities as measured across the diagonals of the bridge are quite small, namely, $120\mu\mu F$, this representing the total capacity and thus requiring $4\mu H$ to tune the amplifier to the shortest wavelength, namely 13 metres.

The grid tuned-circuit is behind the screening box shown in Fig. 402, which is cut away only to allow the grid input circuit

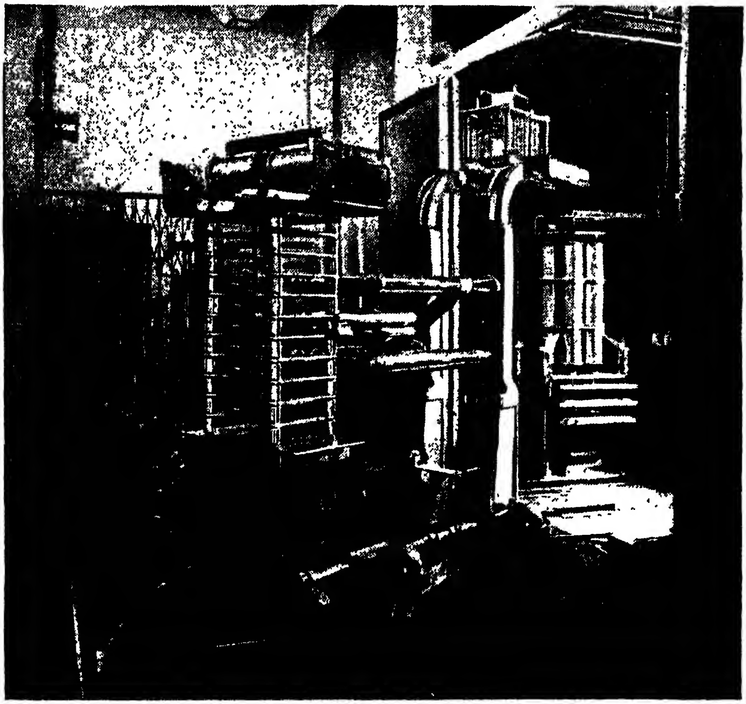


Fig. 403. Main H.F. Circuits—100 kW Transmitter.

to make connection to the grid seal of the valve, which takes the form of a complete copper ring having an ample current-carrying capacity for the large H.F. current passing to the valve grid. The valve and condenser assembly is mounted on a box-shaped insulating pedestal of mycalex.

The design of the grid and anode truck-mounted circuits is seen from Fig. 403 which shows the 16-metre truck. The anode inductance consists of a group of horizontal tubes,

(four only in this case) cross-connected by tubes which can be slid to the position required for tuning, supported from the vertical busbars by mycalex insulation where necessary, the ends of the inductance being connected directly to the busbars, these latter being mounted on pedestals of mycalex. Between the turns of the anode inductance will be seen the output coupling inductance bolted to the tuning condenser which is supported on a separate pedestal. Examination of the layout will indicate what a small amount of insulating material lies within the H.F. field.

The grid-circuit tuning inductance is carried on two pairs of curved mycalex supports above the anode-circuit busbars, and in Fig. 403 will be seen the short length of cross-over arms from the balancing condensers rising vertically through the central vertical column.

This transmitter can be built for series modulation or Class B push-pull modulation. Both systems have been described in Chapter XIII, but whereas series modulation employing the cathode-follower circuit is almost free from distortion as it is a 100% negative feed-back circuit, it is usual to incorporate a negative feed-back circuit when a Class B modulator is used.

In the case of a series-modulated circuit, since the modulators operate at high power and in opposite phase to the H.F. amplifiers, the loading on the main power supply is constant, and in consequence there will not be any tendency to scintillate during modulation. But with a Class B modulator, rapid fluctuations of load will be experienced at modulation frequencies and in consequence scintillation will occur unless precautions are taken to avoid it. These involve the use of filters in the anode circuits of the modulators to smooth out the A.C. components of the speech, and the use of low-reactance supply circuits.

G.E.C. 1 kW Frequency-Modulated Transmitter

This transmitter is intended for high-fidelity broadcasting at frequencies up to 100 Mc/s. At all modulation frequencies a frequency deviation of 75 kc/s represents 100% modulation, this deviation remaining the same whatever the frequency being radiated. Both pre-emphasis or a level characteristic can be

obtained and with the latter the characteristic is flat to within ½ db between 30 c/s and 15 kc/s. We will describe a few of its special features.

The frequency modulation is applied to the master oscillator (M.O.), the circuit (see Fig. 404) being rather more elaborate than that given on page 544. The oscillator utilises the grid and screen grid of the pentode V_0, whilst the anode circuit is tuned to twice the oscillator frequency, so that this valve acts also as the first frequency doubler.

Suppose that the transmitter is set up to radiate a carrier of

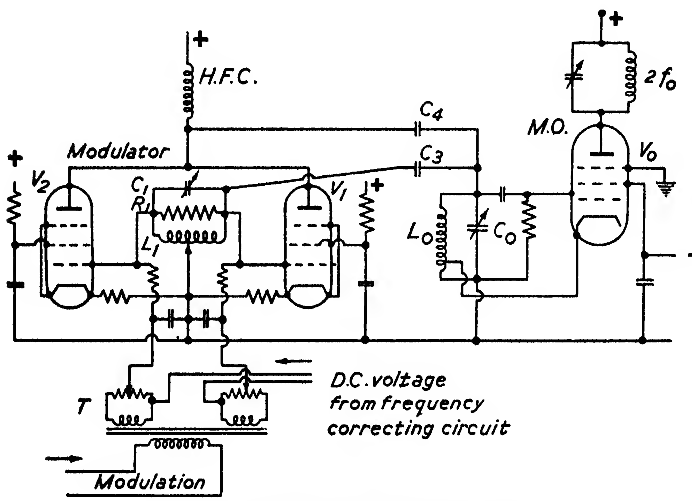


FIG. 404. M.O. and Modulator Circuit.

90 Mc/s, then the frequency of M.O. would be set at 90/16, that is, 5·625 Mc/s and the deviation of this frequency would be 75/16 or 4·69 kc/s for 100% modulation.

A pair of pentodes, V_1 and V_2, are used as reactance modulators, having a push-pull input and parallel output, the latter being coupled across the tuned circuit of M.O. by C_4. The push-pull input to the valves is derived from three circuits: the modulation through the transformer T; a D.C. from the frequency-correction circuit (to be described later); an R.F. voltage from the oscillator, by way of C_3 and the highly-damped resonant circuit $L_1C_1R_1$. C_3 has a high reactance compared with the effective resistance of $L_1C_1R_1$, so that any R.F.

voltage across the latter will be in quadrature with the oscillator voltage. It is clear that the first two circuits we have mentioned will cause low-frequency currents through V_1 and V_2, but $L_1C_1R_1$ will produce R.F. currents.

The action of the circuit is as follows: Let us assume V_1 and V_2 are balanced and we have no D.C. input from the frequency-correcting circuit and no modulation. Then $L_1C_1R_1$ can be adjusted, by the tapping on L_1 and the tuning of C_1, so that the grids of V_1 and V_2 are supplied with equal R.F.

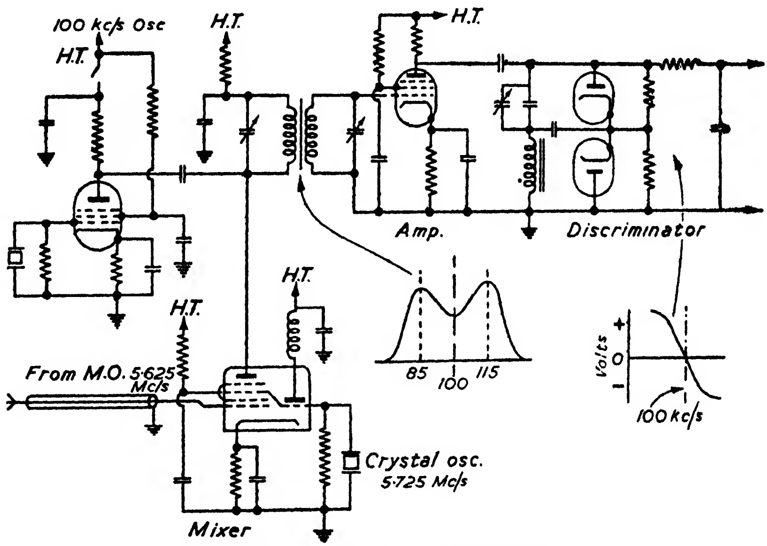


FIG. 405. Frequency-correction Circuit.

voltages, that on V_1 leading 90° on the oscillator voltage and that on V_2 lagging 90°. The sum of the resulting anode currents will then be zero.

If now modulation is applied, then if we consider the half-cycle of modulation during which the modulation is making V_1 less negative, then g_m for this valve will increase and the leading current taken by it will increase. In effect, therefore, the capacitance of the oscillator circuit is increased and the frequency lowered. The grid of V_2, however, will be made more negative and its g_m decreased. As the R.F. current through this valve is lagging, its decrease is equivalent to

increasing a shunt inductance across the oscillator circuit and this, too, will cause the frequency to decrease.

It is important that the mean carrier frequency should be held constant and this is achieved by referring the frequency of M.O. to that of a crystal oscillator, the circuit being shown in Fig. 405.

The crystal oscillator, employing a B.T.-cut crystal, has a frequency which is 100 kc/s different from the correct frequency of M.O. If this is working at its correct frequency, a 100 kc/s beat will evidently come from the mixer and this is applied, after amplification, to a series circuit resonant at 100 kc/s. The voltages across its coil and condenser will be practically the same at resonance and these are applied to diodes, which pass no resultant current under these conditions.

If, however, the frequency of M.O. changes, then there will be an output from the diodes, the direction of the current depending upon whether the frequency has gone up or down. This output is applied to the reactance valves (Fig. 404) and therefore corrects the carrier frequency.

When M.O. is being modulated, its frequency is, of course, changing all the time and it is its mean frequency which we wish to hold constant. The diode circuit therefore includes a resistance-capacitance combination of sufficient time constant to prevent the frequency-correction circuit following the modulation.

A 100 kc/s crystal oscillator can be switched in to line up the discriminator circuit, so that there is zero output from the diodes (indicated by a centre-zero instrument) when the frequency of M.O. is correct.

The 100 kc/s beat frequency is frequency modulated but it is found that a discriminator employing the simple series circuit is quite satisfactory. It is, however, necessary to remove harmonics of the 100 kc/s since these are also modulated (with twice the deviation) and may produce sidebands which fall within the acceptance band of the discriminator and affect its output.

It is clearly very necessary that the resonant circuit in the discriminator should be very stable since the stability of M.O. depends upon this as well as on the stability of the crystal oscillator. Since, however, this circuit is resonant to such a much lower frequency than that of M.O., a given percentage

change in its frequency would lead to a very much smaller percentage change in the frequency of M.O.

In the actual transmitter, all the above stages are in the

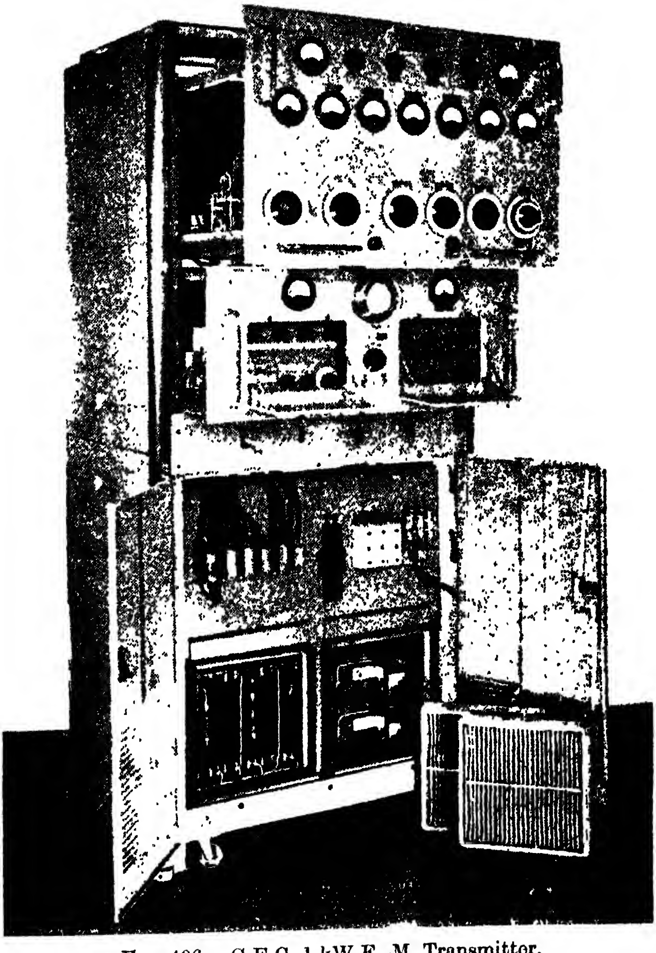


FIG. 406. G.E.C. 1 kW F.-M. Transmitter.

Driver Unit (Fig. 406), together with two stages of frequency-doubling and a C.R.O. monitoring circuit. The latter includes a discriminator circuit to convert the modulation to amplitude modulation and enable it to be examined on the C.R.O.

The output of the Driver Unit which, in our example, would have a frequency of $5{\cdot}625 \times 4 = 22{\cdot}500$ Mc/s with a deviation of $4 \times 4{\cdot}69 = 18{\cdot}76$ kc/s for 100% modulation, is now passed to the Power Amplifier Unit (the upper compartment).

The valves employed in the P.A. Unit are all ACT19's, which are triodes with very low inter-electrode capacitance, the grid connection being at the opposite end to the filament. These valves are mounted in a ceramic tube through which air

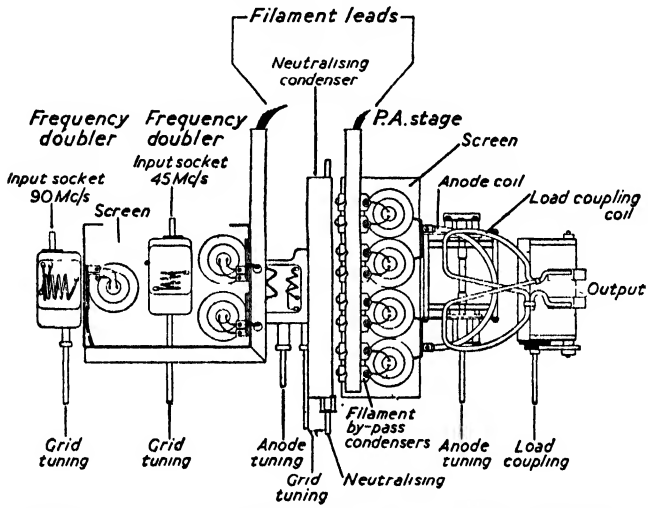


*Fig. 407. Lay-out of Power Amplifier—F.-M. Transmitter.

is blown. The air is drawn in at the bottom of the transmitter and also serves to cool the power unit.

The first stage employs one valve as a frequency doubler, whilst the next stage is also a frequency doubler, with two valves in push-pull. The final stage employs four valves, two on each side of a push-pull circuit. Even at 90 Mc/s the final stage gives a power gain of about 10 and has a conversion efficiency of some 60%.

A plan view of the P.A. layout is shown in Fig. 407, and it will be seen that, instead of each pair of valves being fixed directly to each side of the tuned circuit, the valves are all mounted in line, an anode loop being connected to points between each pair. A second point of interest is that a con-

* Reproduced, by permission, from Lovell Foot, I kW V.H.F. F-M Transmitters, British Institute of Radio Engineers, May 1947.

ventional lumped *LC* circuit is used, rather than a line, even at 100 Mc/s. The lumped circuit has the merit of requiring less space, the adjustment is simpler and it is easier to couple to the output. Either a 75 or a 300Ω feeder can be used.

In order that all the valves in the final stage should be loaded equally, separate grid-leak biassing is provided for each valve. Sufficient cathode biassing is also provided to keep the anode dissipation within safe limits if the drive should fail, but this is normally shorted out by a relay working on grid current.

Provision is made for cutting out one stage of frequency doubling, so that the radiated frequency is halved. It is then necessary to double the deviation applied to M.O., if the 75 kc/s deviation in the output is to be maintained.

It will be seen from the illustration that the Driver and P.A. Units draw out for inspection. The lowest compartment contains a three-phase rectifier.

G.E.C. V.H.F. F-M Mobile Transmitter/Receiver

This equipment is designed to provide telephonic communication between mobile units, such as police cars, and headquarters, where a more powerful transmitter (usually 100 W)

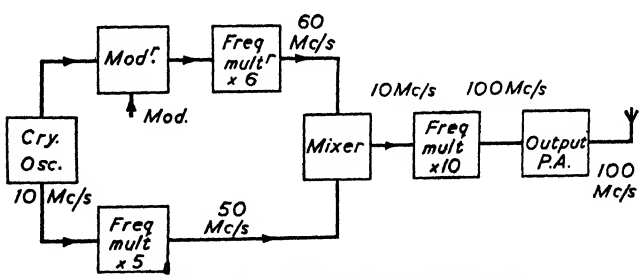


FIG. 408. G.E.C. Mobile F.-M. Transmitter.

is employed. Transmission normally takes place on one predetermined frequency and equipment can be provided covering the range 30–170 Mc/s. It is possible to switch to another frequency separated from the first by 400 kc/s. The transmitter output is 20 W at 100 Mc/s and not less than 10 W at 170 Mc/s.

By the use of miniature valves, etc., the transmitter, a double-superheterodyne receiver, and the power supplies are

contained in one unit measuring 18 × 18 × 8 inches and weighing 35 lb.

A block diagram of the transmitter is shown in Fig. 408, set up to radiate a carrier frequency of 100 Mc/s. It will be seen that the carrier frequency is determined by a crystal oscillator working at 10 Mc/s, the crystal being in a thermostatically-controlled oven. If it is desired to have the use of a second frequency channel, 400 kc/s away from the first, then a second crystal is fitted (selected by a switch) and all the subsequent circuits have a passband broad enough to pass either frequency.

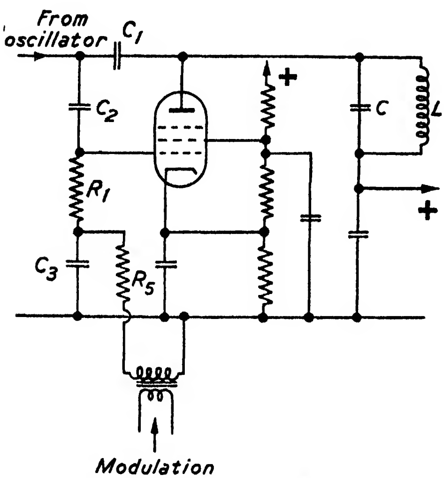


FIG. 409. G.E.C. Mobile F.–M. Transmitter Phase Modulation Circuit.

The desired frequency modulation of the transmitter is derived from phase modulation, using a reactance valve across a resonant circuit tuned to the crystal frequency. R_5 and C_3 ensure that the modulation voltage applied to the grid falls with increasing frequency, as is necessary if the deviation for a modulation of given amplitude is to be the same whatever the modulation frequency.

The deviation of the transmitted carrier is arranged to be 12·5 kc/s for 100% modulation and it is desired to deal with speech frequencies between 300 and 3,000 c/s. It is not possible to employ a phase deviation of more than about one radian without serious distortion and in this transmitter 100% modulation corresponds to about 0·61 radian (35°).

Now $\Delta f = f_m\phi$ and at the modulator Δf, for a 300 c/s modulation and a ϕ of 0·61, is 0·183 kc/s. The effective multiplication of the frequency would therefore require to be 68, but 60 is actually used in this transmitter, so that the deviation for 300 c/s is only 11 kc/s but becomes 12·5 kc/s at about 1,000 c/s.

The way in which this multiplication is achieved will now be described. From Fig. 408 it will be seen that the output of the crystal oscillator is fed into two branches, in one of which it is phase modulated and multiplied by six. This will bring the deviation (for 100% modulation) up to 1·25 kc/s (for modulation frequencies above 1,000 c/s) at a carrier frequency of 60 Mc/s. In the other branch the frequency becomes 50 Mc/s (unmodulated) and the output of the mixer is therefore 10 Mc/s, with a

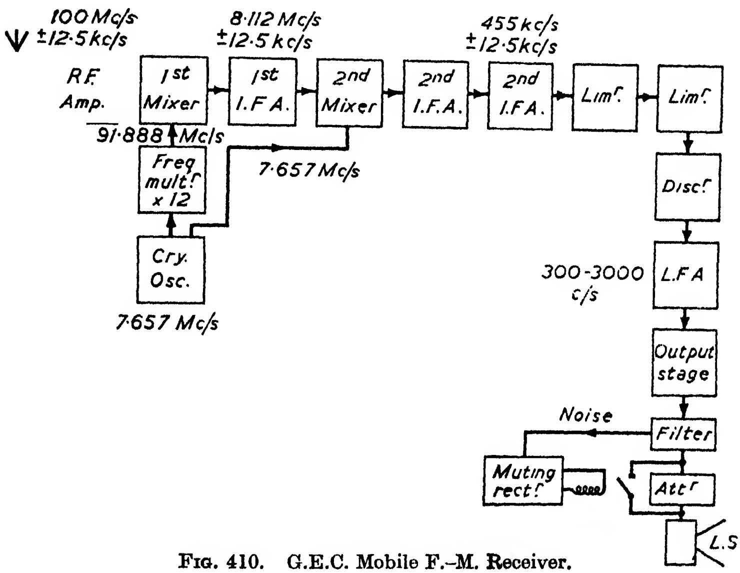


FIG. 410. G.E.C. Mobile F.-M. Receiver.

deviation of 1·25 kc/s. The succeeding frequency multiplier makes this 100 Mc/s, with a deviation of 12·5 kc/s.

The valves employed in all stages except the output are miniature pentodes (Z77) but the output stage employs a double-pentode (TT15) in push-pull..

The lay-out of the receiver should be evident from the block diagram of Fig. 410. Again miniature pentodes (Z77) are used throughout except for the output stage which is of the quiescent push-pull type and delivers 1·3 W to the loudspeaker.

It will be seen that the frequencies for injection into both mixers are derived from the same crystal oscillator, the frequencies shown being suitable for a 100 Mc/s carrier. As in

the transmitter, a second crystal having a frequency 400 kc/s from that of the first can be switched in. All the resonant circuits remain at the same setting for either channel.

The gain available up to the first limiter is such that this is saturated by the thermal noise of the first stage and hence amplitude modulation is removed from any usable signal. Since no appreciable amplification can be provided at the original frequency, the large amplification required has to be obtained in the I.F. stages and the use of two I.F. frequencies enables this to be done without risk of instability.

The detectors following the discriminator circuit are small selenium rectifiers. The output stage is followed by a filter which separates frequencies above 3,000 c/s from the loud-speaker circuit. These high-frequency noise components, after being rectified and amplified, operate a relay which inserts attenuation into the loud-speaker circuit. Hence when no signal is being received the loud-speaker is partly muted, enough noise output remaining to make it evident that the receiver is working. When a carrier is received, the noise components are reduced and the relay cuts out the attenuation.

The trans-receiver is made in two types, to run off either 6 V or 12 V batteries. Two rotary converters are fitted in the case, one delivering about 150 V D.C. to the receiver and earlier transmitter stages and the other supplying about 320 V D.C. to the later transmitter stages.

When installing the equipment in a car it is usual to mount it in the luggage boot and remote control is provided on the dashboard. This trans-receiver also provides a "public address" facility. A loud-speaker is fitted to the roof of the car and the microphone can be connected to this instead of to the radio transmitter. The output valve of the transmitter becomes the low-frequency output valve, supplying 8 W to the loud-speaker.

Automatic Selection of Channels

The extensive use of U.S.W. networks for maintaining communications between a number of mobile stations has made it desirable to be able to select one of a number of operating frequencies. The frequencies of transmitters and receivers

must be very accurately set if there is to be no delay or uncertainty in making contact, particularly as the equipment is usually used by persons having other duties and not by skilled operators. An obvious method would be to supply separate crystals and frequency-multiplying circuits for each frequency, but this method would clearly be too costly and bulky to be considered if many channels are required.

Trans-receivers have been developed in which any one of 330 frequency channels in the U.S.W. band can be selected at the turn of a knob, the frequency of the transmitter and of the

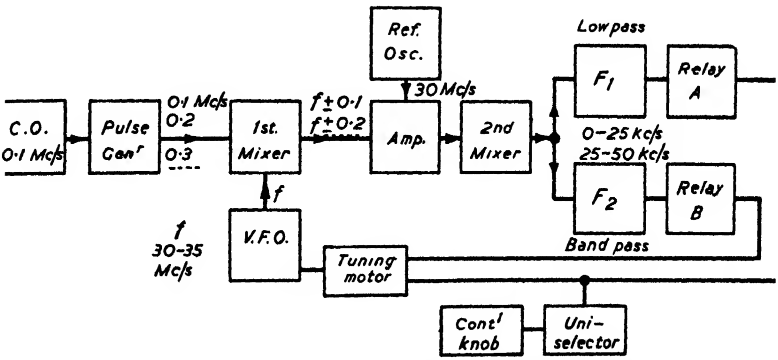


FIG. 411. Circuit for Automatic Frequency Selection.

receiver local oscillator being set accurately by reference to a crystal.

The functions of the selecting equipment can be considered under three heads :—

(*a*) To produce a spectrum of harmonics of an accurately known frequency.

(*b*) To mix these with the output of a variable frequency oscillator and by comparing the side frequencies with a fixed accurate reference frequency, produce control voltages which will enable any desired number of harmonics to be counted.

(*c*) When the desired number of harmonics has been counted (i.e. the V.F.O. has been set to the desired channel) an A.F.C. circuit shall be switched in which will hold the frequency of the V.F.O. to within a few kc/s of its nominal value, determined by the harmonic chosen and the reference frequency oscillator.

The elements of such a system, developed by the G.E.C. Ltd.,

will now be briefly described. A 100 kc/s crystal oscillator triggers a pulse circuit, so that very short pulses are produced every 10 μ seconds. The frequency spectrum of such a pulse will contain all the harmonics of the 100 kc/s and the amplitude of the harmonics will be sensibly equal up to about 6 Mc/s or so (see page 24).

These harmonics are fed into a mixer valve together with a voltage from the variable frequency oscillator which is to be controlled. As the V.F.O. sweeps over its frequency range, say 30 to 35 Mc/s (the frequency will be multiplied up to produce the desired U.S.W. frequency) driven by an electric motor through a gearbox, the " comb " of side frequencies on either side of the V.F.O. frequency moves past a fixed crystal-controlled reference frequency, 30 Mc/s, say.

Beat notes between this fixed frequency and the harmonic spectrum are produced in a second mixer valve and then passed through two filters, one Low-Pass, 0–25 kc/s (F_1) and one Band-Pass, 25–50 kc/s (F_2). The outputs from the filters are rectified and operate relays A and B.

In order to select a channel, a contact is chosen on a uni-selector (as used in automatic telephony) by means of a control knob. The tuning motor then moves the V.F.O. over the band, starting at the low-frequency end, and each time a beat note is produced which passes through F_1, say, the relay moves the uni-selector on one contact until the contact is reached which has been selected by the control knob.

When the selected contact has been reached the motor is stopped and an A.F.C. circuit is switched in. Under this condition it is arranged that frequencies passing through F_1 cause the tuning motor to drive the tuning condenser in one direction and the reverse way for frequencies passing through F_2. For example, suppose the V.F.O. has been set so that its frequency is 32·22 Mc/s, a 20 kc/s beat note will pass through F_1 and cause the motor to increase the frequency of the V.F.O. until the beat note is 25 kc/s, in which case both relays A and B will be operated and the motor stopped. Any drift of the oscillator will also be taken up by the A.F.C. circuit in this way.

The action takes some time to describe but the operation is completed in less than 5 seconds. The circuit here described

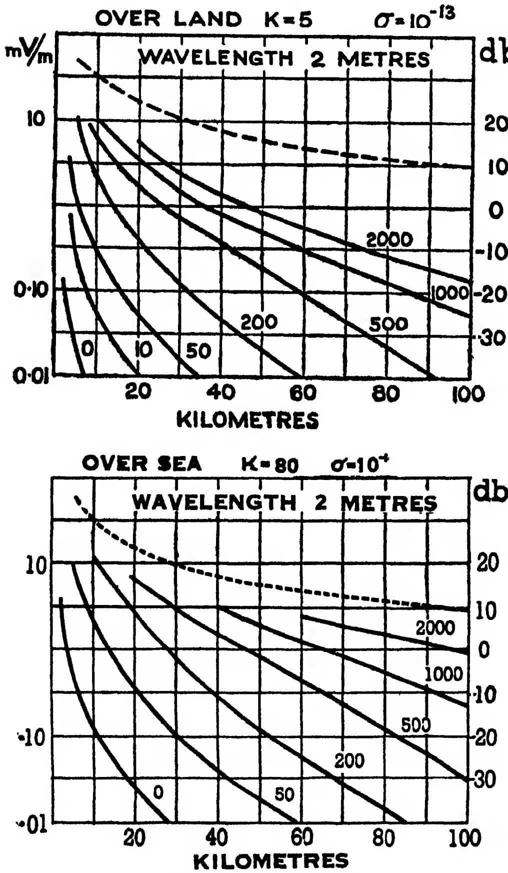
is suitable for selecting about 50 channels and further elaboration is required to deal with 330 channels mentioned earlier.

It will be appreciated that it has been necessary to deal with the system in a simplified and outline fashion, in order to keep the description brief.

PROPAGATION CURVES FOR WAVELENGTHS OF 2, 4 AND 6 METRES

FIGURES 412, 413 and 414 show curves from which the field strength due to a transmitter on the ground (using a wavelength of 2, 4 or 6 metres) at a receiver elevated at various heights can be obtained, both when transmission is over earth (of average σ and κ) or over sea. These heights (in metres) are indicated against each curve, values being given from 0 to 2,000 metres height. All curves are for 1 kW radiated from a half wave aerial, and for values σ and κ as indicated on the diagrams. Observe the greatly increased range obtained on all wavelengths over sea, particularly when the aerials are both at zero height.

The set of curves given here is for shorter distances



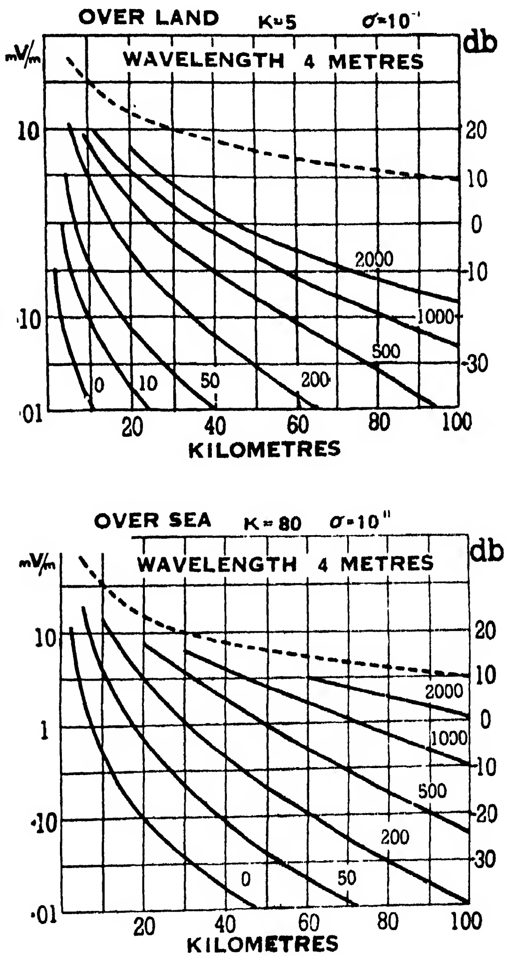
CURVES FOR 1 kW RADIATED.

FIGURES BY CURVES INDICATE HEIGHT IN METRES OF TRANSMITTER (OR RECEIVER).

FIG. 412.

than the sample curves given in the main text, as it is considered that these distances are of more interest to the average experimentalist.

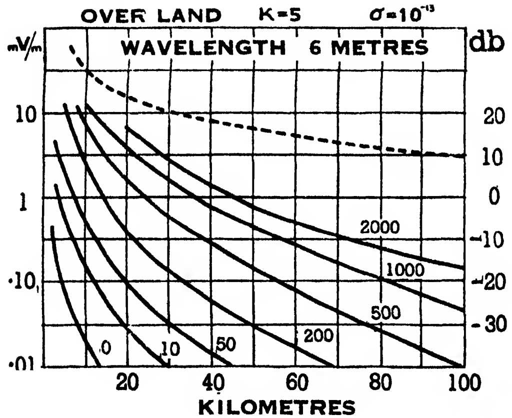
When both transmitter and receiver are elevated above earth, the gain in field strength can be obtained by using the supplementary curve of Fig. 415, or Fig. 61, page 106, in the manner explained on page 107. It should be noted, however, that when the height considered is below the critical values of h_0, the earth's constants have considerable effect and the appropriate curve given on Fig. 415 should be used. For great heights the curve on page 106 should be used. In both cases it is necessary to apply the correction figures given below the curves for sea and land.

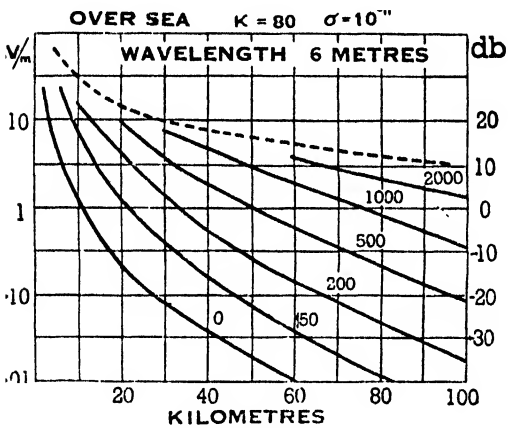


Curves by 1 kW Radiated.

FIGURES BY CURVES INDICATE HEIGHT IN METRES OF TRANSMITTER (OR RECEIVER).

Fig. 413.





CURVES FOR 1 kW RADIATED.

FIGURES BY CURVES ARE HEIGHT IN METRES OF TRANSMITTER (OR RECEIVER).

FIG. 414.

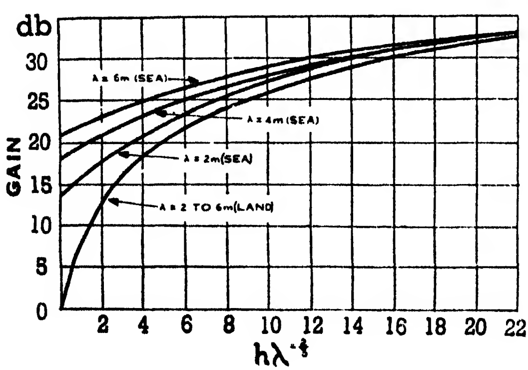


FIG. 415.

The following are the figures in decibels to be subtracted:—

Wavelength (metres)	2	4	6
db to be subtracted			
Over Land	0	2·2	3·7
Over-Sea	13·6	18 0	20·8

OPTIMUM WAVELENGTH FOR COMMUNICATION OVER DISTANCES BETWEEN 500 AND 5,000 KILOMETRES

THE following curves are suitable only for distances between the limits mentioned above, and give only an approximate result, since they average the conditions over any given route.

Daylight Conditions

When daylight spreads over the route between the stations being considered, both bending and attenuation have to be

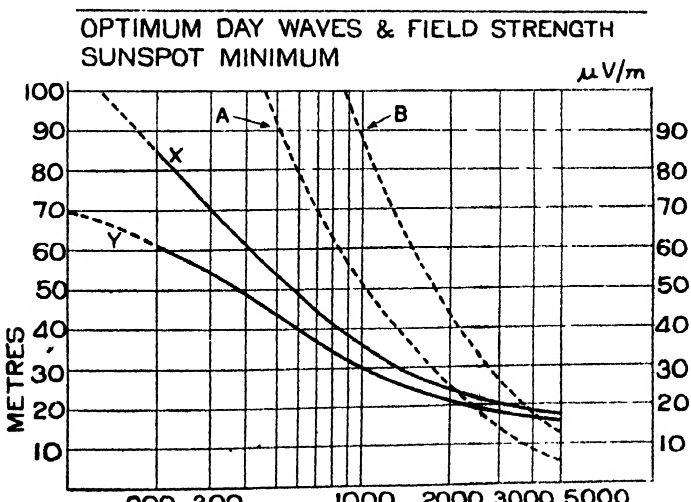


FIG. 416.

considered. Since the attenuation is least on the shorter wavelengths, the optimum wavelength will be one which is just above the skip wavelength for the distance considered. For any given distance, the day wavelength limits will be found between the curves *XY*, Fig. 416. These curves apply to routes

across temperate latitudes and during the sunspot minimum years. For tropical routes at all times and over routes in temperate latitudes during sunspot maximum periods, the wavelength selected from Fig. 416 should be reduced by 20%.

Having obtained the optimum wavelength for the distance required, the field strength for 1 kW radiated can be found by reference to the curves *A* and *B*, where these letters have the same significance as in the main text, namely sunlight and weak sunlight.

Night Conditions

For communications over routes in darkness, the attenuation on all wavelengths is small, and the conditioning feature is bending, which is small. Fig. 417 shows a group of limiting wavelength contours (i.e. for the longest distance) for different seasons, latitudes, and local times after sunset, both latitude and time being considered at a point midway between the stations between which communication is desired. The mid-point time is found from the difference of longitude, namely 1° of longitude difference is equal to four minutes of time (remembering that longitude east, Greenwich time least) and the difference of latitude must be estimated knowing the latitudes of the stations concerned and their distance apart.

The time after sunset for the midpoint of the route can be found from Table XXV which correlates these local sunset times for the different seasons and latitudes.

The curves of Fig. 417 enable us to find the wavelength above which the bending is insufficient to return a ray to earth over the longest possible distance. Having found this limiting wavelength, or what is virtually the maximum electron density at the centre part of the route, reference can then be made to the skip curves of Fig. 76, page 137, to determine the minimum wavelength for the particular distance considered. This wavelength will of course be greater since the contour lines of Fig. 417 represent the limiting conditions for the most oblique angle of incidence.

Having selected a wavelength that satisfies the boundary conditions, the probable field strength can be ascertained from Table XXVI, the field strength being approximately the same for all wavelengths between 25 and 50 metres. This is so, because

TABLE XXV. *Local Sunset Times (24 hours) at Various Latitudes*

Northern Hemisphere.	LATITUDE							Southern Hemisphere.
	0°	10°	20°	30°	40°	50°	60°	
Jan. 1—Dec. 12	1800	1745	1730	1700	1630	1600	1445	June 10—July 3
Jan. 21—Nov 22	1800	1745	1730	1715	1645	1615	1515	May 21—July 24
Feb. 8—Nov. 3	1800	1745	1730	1715	1700	1645	1600	May 1—Aug. 12
Feb. 17—Oct. 26	1800	1745	1745	1730	1715	1700	1630	April 22—Aug. 22
Feb. 28—Oct. 14	1800	1800	1745	1745	1730	1730	1700	April 11—Sept. 2
Mar. 13—Oct. 1	1800	1800	1800	1800	1745	1745	1730	April 1—Sept. 15
Mar. 21—Sept. 24	1800	1800	1800	1800	1800	1800	1800	Mar. 21—Sept. 24
April 1—Sept. 15	1800	1800	1800	1800	1815	1815	1830	Mar. 13—Oct. 1
April 11—Sept. 2	1800	1800	1815	1815	1830	1845	1900	Feb. 28—Oct. 14
April 22—Aug. 22	1800	1800	1815	1830	1845	1900	1930	Feb. 17—Oct. 26
May 1—Aug. 12	1800	1815	1830	1845	1900	1915	2000	Feb. 8—Nov. 3
May 21—July 24	1800	1815	1830	1845	1915	1945	2045	Jan. 21—Nov. 22
June 10—July 3	1800	1815	1830	1900	1930	2000	2115	Jan. 1—Dec. 12

it is the bending of the ray alone which determines the communication. The received signal will, in general, be fading and the field strengths given in Table XXVI are the highest values to which the field strength is likely to rise. These field strengths are for 1 kW radiated from a half-wave aerial, so that with smaller or greater powers, the necessary correction must be made for this.

TABLE XXVI. *Approximate Night Field Strength values for 1 kW radiated (peak values) for all wavelengths between 25 and 50 metres from half wave aerial.*

Distance. Kilometres.	Micro-volts per metre.
500	1,000
600	500
700	400
800	300
900	260
1,000	225
1,500	155
2,000	108
2,500	84
3,000	67
3,500	56
4,000	47
4,500	41
5,000	36

Example. Select a suitable wavelength for communication between London and Stockholm at 9 p.m. (2100) London, on 10th December. Obtain the field strength for 100 watts radiated from a half-wave aerial.

The mid-latitude is approximately 56° and the distance 1,500 kms. The longitude difference between London and Stockholm is 18° and the midpoint time which is 9° E. of London is 36 minutes later than 9 p.m., say 9·30 p.m. To find the hour this is after sunset, refer to Table XXV. This gives us for a latitude of 56° for 12th December a time of 1500, and thus the mid-latitude time is 6½ hours after sunset.

Referring now to Fig. 417, Winter, we see that the limiting wavelength for latitude 55°, 6½ hours after sunset is 32 metres, and by reference to Fig. 76, for a distance of 1,500 kms, we follow the

32 metre contour, curve 4, which gives for this distance a minimum wavelength for communication of 50 metres.

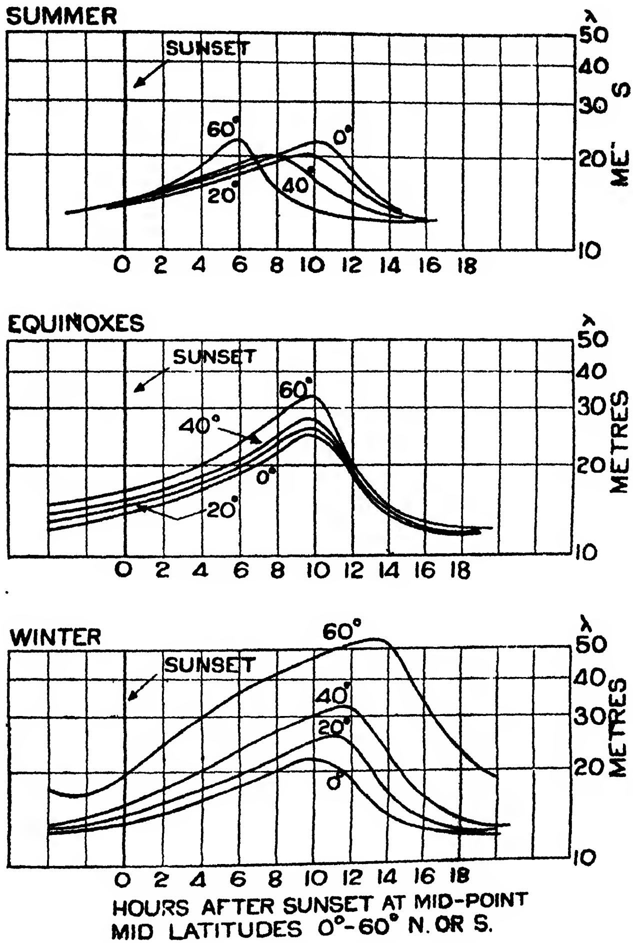


Fig. 417.

From Table XXVI, the peak field strength for 1 kW radiated is seen to be 155 micro-volts per metre, but since the power used is only 100 watts, the field strength will be :

$$\sqrt{\frac{100}{1,000}} \times 155 = 45 \text{ microvolts per metre.}$$

INDEX